TEN PRACTICE TIPS FOR BACKUPS IN CIVIL DISCOVERY

1. **Backup ≠ Inaccessible.** Don’t expect to exclude the content of backups from the scope of discovery if you haven’t laid the foundation to do so. Fed. R. Civ. P. 26(b)(2)(B) requires parties identify sources deemed not reasonably accessible because of undue burden or cost. **Be prepared to prove the cost and burden through reliable metrics and testimony.**

2. **Determine if your client:**
   - Routinely restores backup tapes to, *e.g.*, insure the system is functioning properly or as a service to those who have mistakenly deleted files;
   - Restored the backup tapes other matters or uses them as an archive;
   - Has the system capacity and in house expertise to restore the data;
   - Has the capability to search the tapes for responsive data?

3. **Don’t blindly pull tapes for preservation.** Backup tapes don’t exist in a vacuum but as part of an information system. A properly managed system incorporates labeling, logging and tracking of tapes, permitting reliable judgments to be made about what’s on particular tapes insofar as tying contents to business units, custodians, machines, data sets and intervals. It’s costly to have to process tapes just to establish their contents. **Always preserve associated backup catalogues when you preserve tapes.**

4. **Be prepared to put forward a sensible sampling protocol in lieu of wholesale restoration.**

5. **Test and sample backups to determine if they hold responsive, material and unique ESI.** Judges are unlikely to force you to restore backup tapes when sensible sampling regiments demonstrate that the effort is likely to yield little of value. Backup tapes are like drilling for oil: **After a few dry holes, it’s time to find a new prospect.**

6. **Be prepared to show that the relevant data on tapes is available from more accessible sources.** Sampling, testing and expert testimony help here.

7. **Know the limits of backup search capabilities.** Most backup tools have search capabilities; however, few of these are up to the task of e-discovery. Can the tool search within all common file types and compressed and container file formats?

8. **Appearances matter!** What would the Judge think if she walked through your client’s tape storage area? Does it look like a dumping ground?

9. **If using a cloud-based backup system, consider bringing your e-discovery tools to the data in the Cloud instead of spending days getting the data out.**

10. **Backup tape is for disaster recovery.** If it’s too stale to use to bring the systems back up, why keep it? **Get rid of it!**
## Appendix 1: Exemplar Backup Tape Log

<table>
<thead>
<tr>
<th>Tape No.</th>
<th>Sess. Id</th>
<th>Host Name</th>
<th>Backup Date/Time</th>
<th>Size in Bytes</th>
<th>Session Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC 001</td>
<td>37</td>
<td>EX1</td>
<td>8/1/2007 6:15</td>
<td>50,675,122,176</td>
<td>Exchange 200x</td>
</tr>
<tr>
<td>ABC 001</td>
<td>38</td>
<td>EX1</td>
<td>8/1/2007 8:28</td>
<td>337,707,008</td>
<td>System state</td>
</tr>
<tr>
<td>ABC 001</td>
<td>39</td>
<td>MGT1</td>
<td>8/1/2007 8:29</td>
<td>6,214,713,344</td>
<td>files incremental or differential</td>
</tr>
<tr>
<td>ABC 001</td>
<td>40</td>
<td>MGT1</td>
<td>8/1/2007 8:45</td>
<td>5,576,392,704</td>
<td>SQL Database Backup</td>
</tr>
<tr>
<td>ABC 001</td>
<td>41</td>
<td>SQL1</td>
<td>8/1/2007 8:58</td>
<td>10,004,201,472</td>
<td>files incremental or differential</td>
</tr>
<tr>
<td>ABC 001</td>
<td>42</td>
<td>SQL1</td>
<td>8/1/2007 9:30</td>
<td>8,268,939,264</td>
<td>SQL Database Backup</td>
</tr>
<tr>
<td>ABC 001</td>
<td>43</td>
<td>SQL1</td>
<td>8/1/2007 9:52</td>
<td>272,826,368</td>
<td>System state</td>
</tr>
<tr>
<td>ABC 005</td>
<td>2</td>
<td>EX1</td>
<td>8/14/2007 18:30</td>
<td>51,735,363,584</td>
<td>Exchange 200x</td>
</tr>
<tr>
<td>ABC 005</td>
<td>3</td>
<td>EX1</td>
<td>8/14/2007 20:35</td>
<td>338,427,904</td>
<td>System state</td>
</tr>
<tr>
<td>ABC 005</td>
<td>4</td>
<td>MGT1</td>
<td>8/14/2007 20:38</td>
<td>6,215,368,704</td>
<td>files incremental or differential</td>
</tr>
<tr>
<td>ABC 005</td>
<td>5</td>
<td>MGT1</td>
<td>8/14/2007 20:53</td>
<td>5,677,776,896</td>
<td>SQL Database Backup</td>
</tr>
<tr>
<td>ABC 005</td>
<td>6</td>
<td>SQL1</td>
<td>8/14/2007 21:06</td>
<td>10,499,260,416</td>
<td>files incremental or differential</td>
</tr>
<tr>
<td>ABC 005</td>
<td>7</td>
<td>SQL1</td>
<td>8/14/2007 21:38</td>
<td>8,322,023,424</td>
<td>SQL Database Backup</td>
</tr>
<tr>
<td>ABC 005</td>
<td>8</td>
<td>SQL1</td>
<td>8/14/2007 21:57</td>
<td>273,022,976</td>
<td>System state</td>
</tr>
<tr>
<td>ABC 002</td>
<td>207</td>
<td>NT1</td>
<td>8/15/2007 20:19</td>
<td>31,051,481,088</td>
<td>loose files</td>
</tr>
<tr>
<td>ABC 002</td>
<td>18</td>
<td>NT1</td>
<td>8/16/2007 8:06</td>
<td>47,087,616,000</td>
<td>loose files</td>
</tr>
<tr>
<td>ABC 014</td>
<td>9</td>
<td>EX1</td>
<td>8/17/2007 6:45</td>
<td>52,449,443,840</td>
<td>Exchange 200x</td>
</tr>
<tr>
<td>ABC 014</td>
<td>10</td>
<td>EX1</td>
<td>8/17/2007 8:53</td>
<td>337,969,152</td>
<td>System state</td>
</tr>
<tr>
<td>ABC 014</td>
<td>11</td>
<td>MGT1</td>
<td>8/17/2007 8:54</td>
<td>6,215,368,704</td>
<td>files incremental or differential</td>
</tr>
<tr>
<td>ABC 014</td>
<td>12</td>
<td>MGT1</td>
<td>8/17/2007 9:09</td>
<td>5,698,748,416</td>
<td>SQL Database Backup</td>
</tr>
<tr>
<td>---------</td>
<td>----</td>
<td>------</td>
<td>----------------</td>
<td>-----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>ABC 014</td>
<td>13</td>
<td>SQL1</td>
<td>8/17/2007 9:22</td>
<td>10,537,009,152</td>
<td>files incremental or differential</td>
</tr>
<tr>
<td>ABC 014</td>
<td>14</td>
<td>SQL1</td>
<td>8/17/2007 9:47</td>
<td>8,300,986,368</td>
<td>SQL Database Backup</td>
</tr>
<tr>
<td>ABC 014</td>
<td>15</td>
<td>SQL1</td>
<td>8/17/2007 10:08</td>
<td>272,629,760</td>
<td>System state</td>
</tr>
<tr>
<td>ABC 003</td>
<td>16</td>
<td>NT1</td>
<td>8/18/2007 6:15</td>
<td>46,850,179,072</td>
<td>loose files</td>
</tr>
<tr>
<td>ABC 003</td>
<td>17</td>
<td>NT1</td>
<td>8/18/2007 9:26</td>
<td>44,976,308,224</td>
<td>loose files</td>
</tr>
<tr>
<td>ABC 004</td>
<td>19</td>
<td>NT1</td>
<td>8/21/2007 6:16</td>
<td>46,901,690,368</td>
<td>loose files</td>
</tr>
<tr>
<td>ABC 004</td>
<td>20</td>
<td>NT1</td>
<td>8/21/2007 9:30</td>
<td>44,742,868,992</td>
<td>loose files</td>
</tr>
<tr>
<td>ABC 009</td>
<td>30</td>
<td>EX1</td>
<td>8/22/2007 8:52</td>
<td>53,680,603,136</td>
<td>Exchange 200x</td>
</tr>
<tr>
<td>ABC 009</td>
<td>31</td>
<td>EX1</td>
<td>8/22/2007 11:01</td>
<td>348,782,592</td>
<td>System state</td>
</tr>
<tr>
<td>ABC 009</td>
<td>32</td>
<td>MGT1</td>
<td>8/22/2007 11:03</td>
<td>6,215,434,240</td>
<td>files incremental or differential</td>
</tr>
<tr>
<td>ABC 009</td>
<td>34</td>
<td>SQL1</td>
<td>8/22/2007 11:31</td>
<td>10,732,371,968</td>
<td>files incremental or differential</td>
</tr>
<tr>
<td>ABC 009</td>
<td>35</td>
<td>SQL1</td>
<td>8/23/2007 4:08</td>
<td>8,362,000,384</td>
<td>SQL Database Backup</td>
</tr>
<tr>
<td>ABC 009</td>
<td>36</td>
<td>SQL1</td>
<td>8/23/2007 4:33</td>
<td>272,629,760</td>
<td>System state</td>
</tr>
<tr>
<td>ABC 011</td>
<td>44</td>
<td>NT1</td>
<td>8/23/2007 6:16</td>
<td>46,938,193,920</td>
<td>loose files</td>
</tr>
<tr>
<td>ABC 011</td>
<td>45</td>
<td>NT1</td>
<td>8/23/2007 9:32</td>
<td>44,611,403,776</td>
<td>loose files</td>
</tr>
</tbody>
</table>
**Databases in E-Discovery**

When I set out to write this chapter on databases in electronic discovery, I went to the literature to learn prevailing thought and ensure I wasn’t treading old ground. What I found surprised me.

I found there’s next to no literature on the topic! What little authority exists makes brief mention of flat file, relational and enterprise databases, notes that discovery from databases is challenging and then flees to other topics. A few commentators mention *In re Ford Motor Co.*, the too-brief 2003 decision reversing a trial court’s order allowing a plaintiff to root around in Ford’s databases with nary a restraint. Although the 11th Circuit cancelled that fishing expedition, they left the door open for a party to gain access to an opponent’s databases on different facts, such as where the producing party fails to meet its discovery obligations.

The constant counsel offered by any article touching on databases in e-discovery is “get help.” That’s good advice, but not always feasible or affordable.

**Because databases run the world, we can’t avoid them in e-discovery.** We have to know enough about how they work to deal with them when the case budget or time constraints make hiring an expert impossible. We need to know how to identify and preserve databases, and we must learn how to gather sufficient information about them to frame and respond to discovery about databases.

**Databases run the world**

You can’t surf the ‘net, place a phone call, swipe your parking access card, use an ATM, charge a meal, buy groceries, secure a driver’s license, book a flight or get admitted to an emergency room without a database making it happen.

Databases touch our lives all day, every day. Our computer operating systems and e-mail applications are databases. The spell checker in our word processor is a database. Google and Yahoo search engines are databases. Westlaw and Lexis, too. Craigslist. Amazon.com. E-Bay. Facebook. All big honkin’ databases.

Yet, when it comes to e-discovery, we tend to fix our attention on documents, without appreciating that most electronic evidence exists only as a flash mob of information assembled

---

58 Happily, since I first published, others have waded in and produced more practical scholarship. Here are links to two recent, thoughtful publications on the topic: 
**Requests for Production of Databases: Documents v. Data**, by Christine Webber and Jeff Kerr (a Georgetown Academy graduate!)
**The Sedona Conference Database Principles Addressing the Preservation & Production of Databases & Database Information in Civil Litigation**

59 345 F.3d 1315 (11th Cir. 2003)
and organized on the fly from a dozen or thousand or million discrete places. In our zeal to lay hands on documents instead of data, we make discovery harder, slower and costlier. Understanding databases and acquiring the skills to peruse and use their contents gets us to the evidence better, faster and cheaper.

Databases are even changing the way we think about discovery. Historically, parties weren’t obliged to *create* documents for production in discovery; instead, you produced what you had on file. Today, documents don’t exist until you generate them. Tickets, bank statements, websites, price lists, phone records and register receipts are all just *ad hoc* reports generated by databases. Documents don’t take tangible form until you print them out, and more and more, only the tiniest fraction of documents—one-tenth of one percent—will emerge as ink on paper, obliging litigants to be adept at both crafting queries to elicit responsive data and mastering ways to interpret and use the data stream that emerges.

**Introduction to Databases**
Most of us use databases with no clue how they work. Take e-mail, for example. Whether you know it or not, each e-mail message you view in Outlook or through your web browser is a report generated by a database query and built of select fields of information culled from a complex dataset. It’s then presented to you in a user-friendly arrangement determined by your e-mail client's capabilities and user settings.

That an e-mail message is not a single, discrete document is confusing to some. The data segments or “fields” that make up an e-mail are formatted with such consistency from application-to-application and appear so similar when we print them out that we mistake e-mail messages for fixed documents. But each is really a customizable report from the database called your e-mail.

When you see a screen or report from a database, you experience an assemblage of information that “feels” like a document, but the data that comes together to create what you see are often drawn from different sources within the database and from different systems, locations and formats, all changing moment to moment.

Understanding databases begins with mastering some simple concepts and a little specialized terminology. Beyond that, the distinction between your e-mail database and Google’s is mostly marked by differences in scale, optimization and security.

**Constructing a Simple Database**
If you needed a way to keep track of the cases on your docket, you’d probably begin with a simple table of columns and rows written on a legal pad. You’d start listing your clients by name. Then,
you might list the names of other parties, the case number, court, judge and trial date. If you still had room, you’d add addresses, phone numbers, settlement demands, insurance carriers, policy numbers, opposing counsel and so on.

In database parlance, you’ve constructed a “table,” and each separate information item you entered (e.g., name, address, court) is called a “field.” The group of items you assembled for each client (probably organized in columns and arranged in a row to the right of each name) is collectively called a “record.” Because the client’s name is the field that governs the contents of each record, it would be termed the “key field.”

Pretty soon, your table would be unwieldy and push beyond the confines of a sheet of paper. If you added a new matter or client to the table and wanted it to stay in alphabetical order by client name, you’d probably have to rewrite the list.

So, you might turn to index cards. Now, each card is a “record” and lists the information (the “fields”) pertinent to each client. It’s easy to add cards for new clients and re-order them by client name. Then, sometimes you’d want to order matters by trial date or court. To do that, you’d either need to extract specific data from each card to compile a report, re-sort the cards, or maintain three sets of differently ordered cards, one by name, one by trial date and a third by court.

Your cards comprise a database of three tables. They are still deemed tables even though you used a card to hold each record instead of a row. One table uses client name as its key field, another uses the trial date and the third uses the court. Each of these three sets of cards is a “flat file database,” distinguished by the characteristic that all the fields and records (the cards) comprise a single file (i.e., each a deck of cards) with no relationships or links between the various records and fields except the table structure (the order of the deck and the order of fields on the cards).

Of course, you’d need to keep all cards up-to-date as dates, phone numbers and addresses change. When a client has more than one matter, you’d have to write all the same client data on multiple cards and update each card, one-by-one, trying not to overlook any card. What a pain!

So, you’d automate, turning first to something like a spreadsheet. Now, you’re not limited by the dimensions of a sheet of paper. When you add a new case, you can insert it anywhere and re-sort the list by name, court or trial date. You’re not bound by the order in which you entered the information, and you can search electronically.
Though faster and easier to use than paper and index cards, your simple spreadsheet is still just a table in a flat file database. You must update every field that holds the same data when that data changes (though “find and replace” functions make this more efficient and reliable), and when you want to add, change or extract information, you have to open and work with the entire table.

What you need is a system that allows a change to one field to update every field in the database with the same information, not only within a single table but across all tables in the database. You need a system that identifies the relationship between common fields of data, updates them when needed and, better still, uses that common relationship to bring together more related information. Think of it as adding rudimentary intelligence to a database, allowing it to “recognize” that records sharing common fields likely relate to common information. Databases that do this are called “relational databases,” and they account for most of the databases used in business today, ranging from simple, inexpensive tools like Microsoft Access or Intuit QuickBooks to enormously complex and costly “enterprise-level” applications marketed by Oracle and SAP.

To be precise, only the tables of data are the “database,” and the software used to create, maintain and interrogate those tables is called the Database Management System or DBMS. In practice, the two terms are often used interchangeably.

Relational Databases
Let’s re-imagine your case management system as a relational database. You’d still have a table listing all clients organized by name. On this CLIENTS table, each client record includes name, address and case number(s). Even if a client has multiple cases in your office, there is still just a single table listing:

<table>
<thead>
<tr>
<th>CLT_LAST</th>
<th>CLT_FIRST</th>
<th>ST_ADD</th>
<th>CITY</th>
<th>STATE</th>
<th>ZIP</th>
<th>CASE_NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ballmer</td>
<td>Steven</td>
<td>3832 Hunts Point Rd.</td>
<td>Hunts Point</td>
<td>WA</td>
<td>98004</td>
<td>001, 005</td>
</tr>
<tr>
<td>Chambers</td>
<td>John</td>
<td>5608 River Way</td>
<td>Buena Park</td>
<td>CA</td>
<td>90621</td>
<td>002</td>
</tr>
<tr>
<td>Dell</td>
<td>Michael</td>
<td>3400 Toro Canyon Rd.</td>
<td>Austin</td>
<td>TX</td>
<td>78746</td>
<td>003, 007</td>
</tr>
<tr>
<td>Ellison</td>
<td>Lawrence</td>
<td>745 Mountain Home Rd.</td>
<td>Woodside</td>
<td>CA</td>
<td>94062</td>
<td>004</td>
</tr>
<tr>
<td>Gates</td>
<td>William</td>
<td>1835 73rd Ave. NE</td>
<td>Medina</td>
<td>WA</td>
<td>98039</td>
<td>001, 005</td>
</tr>
<tr>
<td>Jobs</td>
<td>Steven</td>
<td>460 Mountain Home Rd.</td>
<td>Woodside</td>
<td>CA</td>
<td>94062</td>
<td>006, 009</td>
</tr>
<tr>
<td>Palmisano</td>
<td>Samuel</td>
<td>665 Pequot Ave.</td>
<td>Southport</td>
<td>CT</td>
<td>06890</td>
<td>007</td>
</tr>
</tbody>
</table>

---

60 One of the most important and widely used database applications, MySQL, is open source; so, while great fortunes have been built on relational database tools, the database world is by no means the exclusive province of commercial software vendors.
It’s essential to keep track of cases and upcoming trials, so you create another table called CASES:

<table>
<thead>
<tr>
<th>CASE_NO</th>
<th>TRL_DATE</th>
<th>MATTER</th>
<th>TYPE</th>
<th>COURT</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>2011-02-14</td>
<td>U.S. v. Microsoft</td>
<td>Antitrust</td>
<td>FDDC-1</td>
</tr>
<tr>
<td>002</td>
<td>2012-01-09</td>
<td>EON v Cisco</td>
<td>Patent</td>
<td>FEDTX-2</td>
</tr>
<tr>
<td>003</td>
<td>2011-02-15</td>
<td>In re: Dell</td>
<td>Regulatory</td>
<td>FWDTX-4</td>
</tr>
<tr>
<td>004</td>
<td>2011-05-16</td>
<td>SAP v. Oracle</td>
<td>Conspiracy</td>
<td>FWDC-8</td>
</tr>
<tr>
<td>005</td>
<td>2012-01-09</td>
<td>Microsoft v. Yahoo</td>
<td>Breach of K</td>
<td>FWDCA-8</td>
</tr>
<tr>
<td>006</td>
<td>2010-12-06</td>
<td>Apple v. Adobe</td>
<td>Antitrust</td>
<td>FNDCA-8</td>
</tr>
<tr>
<td>007</td>
<td>2011-10-31</td>
<td>Dell v. Travis County</td>
<td>Tax</td>
<td>TX250</td>
</tr>
<tr>
<td>008</td>
<td>null</td>
<td>Hawkins v. McGee</td>
<td>Med Mal</td>
<td>FUSSC</td>
</tr>
<tr>
<td>009</td>
<td>2011-12-05</td>
<td>Jobs v. City of Woodside</td>
<td>Tax</td>
<td>CASMD09</td>
</tr>
</tbody>
</table>

You also want to stay current on where your cases will be tried and the presiding judge, so you maintain a COURTS table for all the matters on your docket:

<table>
<thead>
<tr>
<th>COURT</th>
<th>JUDGE</th>
<th>FED_ST</th>
<th>JURISDICTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNDCA-8</td>
<td>Laporte</td>
<td>FED</td>
<td>Northern District of California (SF)</td>
</tr>
<tr>
<td>FDDC-1</td>
<td>Kollar-Kotelly</td>
<td>FED</td>
<td>USDC District of Columbia</td>
</tr>
<tr>
<td>FWDTX-4</td>
<td>Sparks</td>
<td>FED</td>
<td>Western District of Texas</td>
</tr>
<tr>
<td>TX250</td>
<td>Dietz</td>
<td>STATE</td>
<td>250th JDS, Travis County, TX</td>
</tr>
<tr>
<td>CASMD09</td>
<td>Parsons</td>
<td>STATE</td>
<td>San Mateo Superior Court, CA</td>
</tr>
<tr>
<td>FEDTX-2</td>
<td>Ward</td>
<td>FED</td>
<td>Eastern District of Texas</td>
</tr>
<tr>
<td>FWDWA-6</td>
<td>Jones</td>
<td>FED</td>
<td>Western District of Washington</td>
</tr>
<tr>
<td>FUSSC</td>
<td>Hand</td>
<td>FED</td>
<td>United States Supreme Court</td>
</tr>
</tbody>
</table>

As we look at these three tables, note that each has a unique key field called the “primary key” for that table.\(^\text{61}\) For the CLIENTS table, the primary key is the client’s last name.\(^\text{62}\) The primary key is the trial date for the TRIAL_DATES table and it’s a unique court identifier for the COURTS table. The essential characteristic of a primary key is that it cannot repeat within the table for

\(^{61}\) Tables can have more than one primary key.

\(^{62}\) In practice, a last name would be a poor choice for a primary key in that names tend not to be unique—certainly a law firm could expect to have multiple clients with the same surname.
which it serves as primary key, and a properly-designed database will prevent a user from creating duplicate primary keys.

Many databases simply assign a unique primary key to each table row, either a number or a non-recurring value built from elements like the first four letters of a name, first three numbers in the address, first five letters in the street name and the Zip code. For example, an assigned key for Steve Ballmer derived from data in the CLIENTS table might be BALL383HUNTS98004. The primary key is used for indexing the table to make it more efficient to search, sort, link and perform other operations on the data.

**Tuples and Attributes**

Now, we need to introduce some new terminology because the world of relational databases has a language all its own. Dealing with the most peculiar term first, the contents of each row in a table is called a “tuple,” defined as an ordered list of elements. In the COURTS table above, there are seven tuples, each consisting of four elements. These elements, ordered as columns, are called “attributes,” and what we’ve called tables in the flat file world are termed “relations” in relational databases. Put another way, a relation is defined as a set of tuples that have the same attributes (See Figure 1).

The magic happens in a relational database when tables are “joined” (much like the cube in Figure 2) by referencing one table from another. This is done by incorporating the primary key in the table referenced as a

---

63 Per Wikipedia, the term “tuple” originated as an abstraction of the sequence: single, double, triple, quadruple, quintuple, sextuple, septuple, octuple…n-tuple. The unique 0-tuple is called the null tuple. A 1-tuple is called a “singleton,” a 2-tuple is a “pair” and a 3-tuple is a “triple” or “triplet.” The n can be any positive integer. For example, a complex number can be represented as a 2-tuple, a quaternion can be represented as a 4-tuple, an octonion can be represented as an octuple (mathematicians use the abbreviation “8-tuple”), and a sedenion can be represented as a 16-tuple. I include this explanation to remind readers why many of us went to law school instead of studying computer science.

64 Although unlike the cube, a relational database is not limited to just three dimensions of attachment.

65 The term “relation” is so confounding here, I will continue to refer to them as tables.
“foreign key” in the referencing table. The table referenced is the “parent table,” and the referencing table is the “child table” in this joining of the two relations. In Figure 3, COURTS is the parent table to CASES with respect to the primary key field, “COURT.” In the CASES table, the foreign key for the field COURT points back to the COURTS table, assuring that the most current data will populate the field. In turn, the CLIENTS table employs a foreign key relating to the CASE_NO attribute in the CASE table, again assuring that the definitive information populates the attribute in the CLIENTS table.

Remember that what you are seeking here is to ensure that you do not build a database with inconsistent data, such as conflicting client addresses. Data conflicts are avoided in relational databases by allowing the parent primary key to serve as the definitive data source. So, by pointing each child table to that definitive parent via the use of foreign keys, you promote so-called “referential integrity” of the database. Remember, also, that while a primary key must be unique to the parent table, it can be used as many times as desired when referenced as a foreign key. As in life, parents can have multiple children, but a child can have but one set of (biological) parents.

Field Properties and Record Structures
When you were writing case data on your index cards, you were unconstrained in terms of the information you included. You could abbreviate, write dates as words or numeric values and include as little or as much data as the space on the card and intelligibility allowed. But for databases to perform properly, the contents of fields should conform to certain constraints to insure data integrity. For example, you wouldn’t want a database to accept four or ten letters in a field reserved for a Zip code. Neither should the database accept duplicate primary keys or open a case without including the name of a client. If a field is designed to store only a U.S. state, then you don’t want it to accept “Zambia” or “female.” You also don’t want it to accept “Noo Yawk.”

Accordingly, databases are built to enforce specified field property requirements. Such properties may include:

1. **Field size**: limiting the number of characters that can populate the field or permitting a variable length entry for memos;
2. **Data type**: text, currency, integer numbers, date/time, e-mail address and masks for phone numbers, Social security numbers, Zip codes, etc.;
3. **Unique fields**: Primary keys must be unique. You typically wouldn’t want to assign the same case number to different matters or two Social Security numbers to the same person.
4. **Group or member lists**: Often fields may only be populated with data from a limited group of options (e.g., U.S. states, salutations, departments and account numbers);

5. **Validation rules**: To promote data integrity, you may want to limit the range of values ascribed to a field to only those that makes sense. A field for a person’s age shouldn’t accept negative values or (so far) values in excess of 125. A time field should not accept “25:00pm” and a date field designed for use by Americans should guard against European date notation. Credit card numbers must conform to specific rules, as must Zip codes and phone numbers; or

6. **Required data**: The absence of certain information may destroy the utility of the record, so certain fields are made mandatory (e.g., a car rental database may require input of a valid driver’s license number).

You’ll appreciate why demanding production of the raw tables in a database may be an untenable approach to e-discovery when you consider how databases store information. When a database populates a table, it’s stored in either **fixed length** or **variable length** fields.

**Fixed-Length Field Records**

Fixed length fields are established when the database is created, and it’s important to appreciate that the data is stored as long sequences of data that may, to the untrained eye, simply flow together in one incomprehensible blob. A fixed length field record may begin with information setting out information concerning all of the fields in the record, such as each field’s name (e.g., COURT), followed by its data type (e.g., alphanumeric), length (7 characters) and format (e.g., only values matching a specified list of courts).

A fixed length field record for a simplified address table might look like Figure 4.

**Figure 4**

<table>
<thead>
<tr>
<th>FIELD=&quot;CIT_IST&quot;</th>
<th>CHAR(20)</th>
<th>FIELD=&quot;CIT_FIRST&quot;</th>
<th>CHAR (20)</th>
<th>FIELD=&quot;ST_ADDO&quot;</th>
<th>CHAR(40)</th>
<th>FIELD=&quot;CITY&quot;</th>
<th>CHAR (20)</th>
<th>FIELD=&quot;STATE&quot;</th>
<th>CHAR(2)</th>
<th>FIELD=&quot;ZIP&quot;</th>
<th>CHAR(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bellmer________Steven_______3032 Hunt e Point Rd Hunts Point WA98044Chambers__________</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>John__________5600 River Way________Duen a Park________CA90521Dell________Michael <strong><strong><strong>3400 Toro Canyon Rd</strong></strong></strong>__</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austin________TX7746Ellison________ Lawrence________745 Mountain Home RdWoodside________CA94062Gates________William ____<strong>1835 73rd Ave. NE________Medina______VA9 <strong><strong><strong>8039Jobs________Steven________460 N________Mounta in Home RdWoodside________CA94062Palinisco________CA94062Palinisco</strong></strong></strong></strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>__________665 Pequot Ave________Southport________CT06890</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Note how the data is one continuous stream. The name, order and length of data allocated for each field is defined at the beginning of the string in all those “FIELD=” and CHAR(x) statements, such that the total length of each record is 107 characters. To find a given record in a table, the database software simply starts accessing data for that record at a distance (also called an “offset”) from the start of the table equal to the number of records times the total length allocated to each record. So, as shown in Figure 5, the fourth record starts 428 characters from the start of the first record. In turn, each field in the record starts a fixed number of characters from the start of the record. If you wanted to extract Steve Jobs’ Zip code from the exemplar table, the Jobs address record is the 6th record, so it starts 642 characters (or bytes) from the start of the first record and the Zip code field begins 102 characters from the start of the sixth record (20+20+40+20+2), or 744 bytes from the start of the first record. This sort of offset retrieval is tedious for humans, but it’s a cinch for computers.

**Figure 5**

Variable-Length Field Records

One need only recall the anxiety over the Y2K threat to appreciate why fixed length field records can be problematic. Sometimes, the space allocated to a field proves insufficient in unanticipated ways, or you may simply need to offer the ability to expand the size of a record on-the-fly. Databases employ variable length field records whose size can change from one record to the next. Variable length fields employ **pointer fields** that seamlessly redirect data retrieval to a designated point in the memo file where the variable length field data begins (or continues). The database software then reads from the memo file until it encounters an end-of-file marker or another pointer to a memo location holding further data.
Forms, Reports and Query Language

Now that you’ve glimpsed the ugly guts of database tables, you can appreciate why databases employ database management software to enter, update and retrieve data. Though DBMS software serves many purposes geared to indexing, optimizing and protecting data, the most familiar role of DBMS software is as a user interface for forms and reports.

There’s little difference between forms and reports except that we tend to call the interface used to input and modify data a “form” and the interface to extract data a “report.” Both are simply user-friendly ways to implement commands in “query languages.”

Query language is the term applied to the set of commands used to retrieve information from a database. The best known and most widely used of these is called SQL (for Structured Query Language, officially ‘ess-cue-ell,’ but most everyone calls it “sequel”). SQL is a computer language, but different from computer languages like Java or C++ that can be used to construct applications, SQL’s sole purpose is the creation, management and interrogation of databases.

Though the moniker “query language” might lead anyone to believe that its raison d’être is to get data out of databases, in fact, SQL handles the heavy lifting of database creation and data insertion, too. SQL includes subset command sets for data control (DCL), data manipulation (DML) and data definition (DDL). SQL syntax is beyond the scope of this paper, but the following snippet of code will give you a sense of how SQL is used to create a table like the case management tables discussed above:

```sql
CREATE TABLE COURTS
  (COURT varchar(7), PRIMARY KEY,
   JUDGE varchar(18),
   FED_ST varchar(5),
   JURISDICTION varchar (40));

CREATE TABLE CASES
  (CASE_NO int IDENTITY(1,1) PRIMARY KEY,
   TRL_DATE
   MATTER varchar (60),
   TYPE varchar (40)
   COURT varchar(7));
```

In these few lines, the COURTS and CASES tables are created, named and ordered into various alphanumeric fields of varying specified lengths. Two primary keys are set and one key, CASE_NO, is implemented so as to begin with the number 1 and increment by 1 each time a new case is added to the CASES table.
Who Owns SQL?
I do, so if your firm or clients are using SQL, please have them send gobs of cash to me so I won’t sue them.

In fact, nobody “owns” SQL, but several giant software companies, notably Oracle and Microsoft, have built significant products around SQL and produced their own proprietary dialects of SQL. When you hear someone mention “SQL Server,” they’re talking about a Microsoft product, but Microsoft doesn’t own SQL; it markets a database application that’s compatible with SQL.

SQL has much to commend it, being both simple and powerful; but, even the simplest computer language is too much for the average user. So, databases employ graphical user interfaces (GUIs) to put a friendly face on SQL. When you enter data into a form or run a search, you’re simply triggering a series of pre-programmed SQL commands.

In e-discovery, if the standard reports supported by the database are sufficiently encompassing and precise to retrieve the information sought, great! You’ll have to arrive at a suitable form of production and perhaps wrangle over scope and privilege issues; but, the path to the data is clear.

However, because most companies design their databases for operations not litigation, very often, the standard reporting capabilities won’t be retrieve the types of information required in discovery. In that event, you’ll need more than an SQL doctor on your team; you’ll also need a good x-ray of the databases to be plumbed.

Schemas, Data Dictionaries, System Catalogs, and ERDs,
The famed database administrator, Leo Tolstoy, remarked, “Great databases are all alike, every ordinary database is ordinary in its own way.” Although it’s with tongue-in-cheek that I invoke Tolstoy’s famous observation on happy and unhappy families, it’s apt here and means that you can only assume so much about the structure of an unfamiliar database. After that, you need the manual and a map.

In the lingo of database land, the “map” is the database’s schema, and it’s housed in the system’s data dictionary. It may be the system’s logical schema, detailing how the database is designed in terms of its table structures, attributes, fields, relationships, joins and views. Or, it could be its physical schema, setting out the hardware and software implementation of the database on machines, storage devices and networks. As Tolstoy might have said, “A logical schema explains death; but, it won’t tell you where the bodies are buried.”
Information in a database is mostly gibberish without the metadata that gives it form and function. In an SQL database, the compendium of all that metadata is called the system catalog. In practice, the terms system catalog, schema and data dictionary seem to be used interchangeably—they are all—in essence—databases storing information about the metadata of a database. The most important lesson to derive from this discussion is that there is a map—or one can be easily generated—so get it!

Unlike that elusive Loch Ness monster of e-discovery, the “enterprise data map,” the schemas of databases tend to actually exist and are usually maps; that is, graphical depictions of the database structures. Entity-Relationship Modeling (ERM) is a system and notation used to lay out the conceptual and logical schema of a relational database. The resulting diagrams (akin to flow charts) are called Entity-Relationship Diagrams or ERDs (Figure 6).

Figure 6: ERD of Database Schema
Two Lessons from the Database Trenches

The importance of securing the schema, manuals, data dictionary and ERDs was borne out by my experience serving as Special Master for Electronically Stored Information. In a drug product liability action involving thousands of plaintiffs, I was tasked to expedite discovery from as many as 60 different enterprise databases, each more sprawling and complex than the next. The parties were at loggerheads, and serious sanctions were in the offing.

The plaintiffs insisted the databases would yield important evidence. Importantly, plaintiffs’ team included support personnel technically astute enough to get deeply into the weeds with the systems. Plaintiffs were willing to narrow the scope of their database discovery to eliminate those that were unlikely to be responsive and to narrow the scope of their requests. But, to do that, they’d need to know the systems.

For each system, we faced the same questions:

i. What does the database do?
ii. What is it built on?
iii. What information does it hold?
iv. What content is relevant, responsive and privileged?
v. What forms does it take?
vi. How can it be searched effectively; using what query language?
vii. What are its reporting capabilities?
viii. What form or forms of production will be functional, searchable and cost-effective?

It took a three-step process to turn things around. First, the plaintiffs were required to do their homework, and the defense supplied the curriculum. That is, the defense was required to furnish documentation concerning the databases. First, each system had to be identified. The defense prepared a spreadsheet detailing, *inter alia*:

- Names of systems
- Applications;
- Date range of data;
- Size of database;
- User groups; and
- Available system documentation (including ERDs and data dictionaries).
This enabled plaintiffs to prioritize their demands to the most relevant systems. I directed the defendants to furnish operator’s manuals, schema information and data dictionaries for the most relevant systems.

The second step was ordering that narrowly-focused meet-and-confer sessions be held between technical personnel for both sides. These were conducted by telephone, and the sole topic of each was one or more of the databases. The defense was required to make knowledgeable personnel available for the calls and plaintiffs were required to confine their questions to the nuts-and-bolts of the databases at issue.

When the telephone sessions concluded, Plaintiffs were directed to serve their revised request for production from the database. In most instances, the plaintiffs had learned enough about the databases that they were actually able to propose SQL queries to be run.

This would have been sufficient in most cases, but this case was especially contentious. The final step needed to resolve the database discovery logjam was a meeting in the nature of a mediation over which I would preside. In this proceeding, counsel and technical liaison, joined by the database specialists, would meet face-to-face over two days. We would work through each database and arrive at specific agreements concerning the scope of discovery for each system, searches run, sample sizes employed and timing and form of production. The devil is in the details, and the goal was to nail down every detail.

It took two such sessions, but in the end, disputes over databases largely ceased, the production changed hands smoothly, and the parties could refocus on the merits.

The heroes in this story are the technical personnel who collaborated to share information and find solutions when the lawyers could see only contentions. The lesson: Get the geeks together, and then get out of their way.

Lesson Two
In a recent case where I served as special master, the Court questioned the adequacy of defendants’ search of their databases. The defendants used many databases to run their far-flung operations, ranging from legacy mainframe systems housed in national data centers to homebrew applications cobbled together using Access or Excel. But whether big or small, I found with disturbing regularity that the persons tasked to query the systems for responsive data didn’t know how to use them or lacked the rights needed to access the data they were obliged to search.

The lesson: Never assume that a DBMS query searches all of the potentially responsive records, and never assume that the operator knows what they are doing.
Database systems employ a host of techniques to optimize performance and protect confidentiality. For example

- Older records may be routinely purged from the indices;
- Users may lack the privileges within the system to access all the potentially responsive records;
- Queries may be restricted to regions or business units;
- Tables may not be joined in the particular ways needed to gather the data sought.

Any of these may result in responsive data being missed, even by an apparently competent operator.

Establishing operator competence can be challenging, too. Ask a person tasked with running queries if they have the requisite DBMS privileges required for a comprehensive search, and they’re likely to give you a dirty look and insist they do. In truth, they probably don’t know. What they have are the privileges they need to do their job day-to-day; but those may not be nearly sufficient to elicit all of the responsive information the system can yield.

**How do you preserve a database in e-discovery?**

Talk to even tech-savvy lawyers about preserving databases, and you’ll likely hear how database are gigantic and dynamic or how incomprehensibly risky and disruptive it is to mess with them. The lawyer who responds, “Don’t be ridiculous. We’re not preserving our databases for your lawsuit,” isn’t protecting her client.

Or, opposing counsel may say, “Preserve our databases? Sure, no problem. We back up the databases all the time. We’ll just set aside some tapes.” This agreeable fellow isn’t protecting his client either. When it comes time to search the data on tape, Mr. Congeniality may learn that his client has no ability to restore the data without displacing the server currently in use, and restoration doesn’t come quick or cheap.

What both of these lawyers should have said is, “Let me explain what we have and how it works. Better yet, let’s get our technical advisors together. Then, we’ll try to work out a way to preserve what you really need in a way you can use it. If we can’t agree, I’ll tell you what my client will and won’t do, and you can go to the judge right away, if you think we haven’t done enough.”

Granted, this conversation almost never occurs for a host of reasons. Counsel may have no idea what the client has or how it works. Or the duty to preserve attaches before an opposing counsel emerges. Or counsel believes that cooperation is anathema to zealous advocacy and wants only to scorch the Earth.
In fact, it’s not that daunting to subject most databases to a defensible litigation hold, if you understand how the database works and exert the time and effort required to determine what you’re likely to need preserved.

Databases are dynamic by design, but not all databases change in ways that adversely impact legal hold obligations. Many databases—particularly accounting databases—are accretive in design. That is, they add new data as time goes on, but do not surrender the ability to thoroughly search data that existed in prior periods. For accretive databases, all counsel may need to do is ascertain and insure that historical data isn’t going anywhere for the life of the case.

Creating snapshots of data stores or pulling a full backup set for a relevant period is a sensible backstop to other preservation efforts, as an “if all else fails” insurance policy against spoliation. If the likelihood of a lawsuit materializing is remote or if there is little chance that the tapes preserved will ultimately be subjected to restoration, preservation by only pulling tapes may prove sufficient and economical. But, if a lawsuit is certain and discovery from the database(s) is likely, the better approach is to identify ways to either duplicate and/or segregate the particular dynamic data you’ll need or export it to forms that won’t unduly impair searchability and utility. That is, you want to keep the essential data reasonably accessible and shield it from changes that will impair its relevance and probative value.

If the issue in litigation is temporally sensitive—e.g., wholesale drug pricing in 2010 or reduction in force decisions in 2008—you’ll need to preserve the responsive data before the myriad components from which it’s drawn, and the filters, queries and algorithms that govern how it’s communicated, change. You’ll want to retain the ability to generate the reports that should be reasonably anticipated and not lose that ability because of an alteration in some dynamic element of the reporting process.

**Forms of Production**

In no other corner of e-discovery are litigants quite so much as the dog that caught the car than when dealing with databases. Data from specialized and enterprise databases often don’t play well with off-the-shelf applications; not surprising, considering the horsepower and high cost of the systems tasked to run these big iron applications. Still, there is always a way. Sometimes a requesting party demands a copy of an entire database, often with insufficient consideration of what such a demand might entail were it to succeed. If the database is built in Access or on other simple platforms, it’s feasible to acquire the hardware and software licenses required to duplicate the producing party’s database environment sufficiently to run the application. But, if the data sets are so large as to require massive storage resources or are built on an enterprise-level DBMS like Oracle or SAP, mirroring the environment is almost out of the question. I say “almost” because the emergence of Infrastructure-as-a-Service Cloud computing
options promises to make it possible for mere mortals to acquire enterprise-level computing power for short stints.

A more likely production scenario is to narrow the data set by use of filters and queries, then either export the responsive date to a format that can be analyzed in other applications (e.g., exported as extensible markup language (XML), comma separated values (CSV) or in another delimited file) or run reports (standard or custom) and ensure that the reporting takes a form that, unlike paper printouts, lends itself to electronic search.

Before negotiating a form of production, investigate the capabilities of the DBMS. The database administrator may not have had occasion to undertake a data export and so may have no clue what an application can do much beyond the confines of what it does every day. It’s the rare DBMS that can’t export delimited data. Next, have a proposed form of production in mind and, if possible, be prepared to instruct the DBMS administrator how to secure the reporting or export format you seek.

Remember that the resistance you experience in seeking to export to electronic formats may not come from the opposing party of the DBMS administrator. More often, an insistence on reports being produced as printouts or page images is driven by the needs of opposing counsel. In that instance, it helps to establish that the export is feasible as early as possible.

As with other forms of e-discovery, be careful not to accept production in formats you don’t want because, like-it-or-not, many Court give just one bite at the production apple. If you accept it on a paper or as TIFF images for the sake of expediency, you often close the door on re-production in more useful forms.

Even if the parties can agree upon an electronic form of production, it’s nevertheless a good idea to secure a test export to evaluate before undertaking a high volume export.

Closing Thoughts
When dealing with databases in e-discovery, requesting parties should avoid the trap of “You have it. I want it.” Lawyers who’d never be so foolish as to demand the contents of a file room will blithely insist on production of the “database.” For most, were they to succeed in such a foolish quest, they’d likely find themselves in possession of an obscure collection of inscrutable information they can’t possibly use.

Things aren’t much better on the producing party’s side, where counsel routinely fail to explore databases in e-discovery on the theory that, if a report hasn’t been printed out, it doesn’t have to be created for the litigation. Even when they do acknowledge the duty to search databases,
few counsel appreciate how pervasively embedded databases are in their clients’ businesses, and fewer still possess the skills needed to translate an amorphous request for production into precise, effective queries.

Each is trading on ignorance, and both do their clients a disservice.

But, these are the problems of the past and, increasingly, there’s cause for cautious optimism in how lawyers and litigants approach databases in discovery. Counsel are starting to inquire into the existence and role of databases earlier in the litigation timeline and are coming to appreciate not only how pervasive databases are in modern commerce, but how inescapable it is that they take their place as important sources of discoverable ESI.
More on Databases in Discovery

I loathe the practice of law from forms, but bow to its power. Lawyers love forms; so, to get lawyers to use more efficient and precise prose in their discovery requests, we can’t just harangue them to do it; we’ve “got to put the hay down where the goats can get it.” To that end, here is some language to consider when seeking information about databases and when serving notice of the deposition of corporate designees (e.g., per Rule 30(b)(6) in Federal civil practice or Rule 199(b)(1) of the Texas Rules of Civil Procedure):

For each database or system that holds potentially responsive information, we seek the following information to prepare to question the designated person(s) who, with reasonable particularity, can testify on your behalf about information known to or reasonably available to you concerning:

1. The standard reporting capabilities of the database or system, including the nature, purpose, structure, appearance, format and electronic searchability of the information conveyed within each standard report (or template) that can be generated by the database or system or by any overlay reporting application;

2. The enhanced reporting capabilities of the database or system, including the nature, purpose structure, appearance, format and electronic searchability of the information conveyed within each enhanced or custom report (or template) that can be generated by the database or system or by any overlay reporting application;

3. The flat file and structured export capabilities of each database or system, particularly the ability to export to fielded/delimited or structured formats in a manner that faithfully reflects the content, integrity and functionality of the source data;

4. Other export and reporting capabilities of each database or system (including any overlay reporting application) and how they may or may not be employed to faithfully reflect the content, integrity and functionality of the source data for use in this litigation;

5. The structure of the database or system to the extent necessary to identify data within potentially responsive fields, records and entities, including field and table names, definitions, constraints and relationships, as well as field codes and field code/value translation or lookup tables.

6. The query language, syntax, capabilities and constraints of the database or system (including any overlay reporting application) as they may bear on the ability to identify, extract and export potentially responsive data from each database or system;
7. The user experience and interface, including datasets, functionality and options available for use by persons involved with the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT;

8. The operational history of the database or system to the extent that it may bear on the content, integrity, accuracy, currency or completeness of potentially responsive data;

9. The nature, location and content of any training, user or administrator manuals or guides that address the manner in which the database or system has been administered, queried or its contents reviewed by persons involved with the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT;

10. The nature, location and contents of any schema, schema documentation (such as an entity relationship diagram or data dictionary) or the like for any database or system that may reasonably be expected to contain information relating to the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT;

11. The capacity and use of any database or system to log reports or exports generated by, or queries run against, the database or system where such reports, exports or queries may bear on the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT;

12. The identity and roles of current or former employees or contractors serving as database or system administrators for databases or systems that may reasonably be expected to contain (or have contained) information relating to the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT; and

13. The cost, burden, complexity, facility and ease with which the information within databases and systems holding potentially responsive data relating to the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT; may be identified, preserved, searched, extracted and produced in a manner that faithfully reflects the content, integrity and functionality of the source data.

Yes, this is the dread “discovery about discovery;” but, it’s a necessary precursor to devising query and production strategies for databases. If you don’t know what the database holds or the ways in which relevant and responsive data can be extracted, you are at the mercy of opponents who will give you data in unusable forms or give you nothing at all.
Remember, these are not magic words. I just made them up, and there’s plenty of room for improvement. If you borrow this language, please take time to understand it, and particularly strive to know why you are asking for what you demand. Supplying the information requires effort that should be expended in support of a genuine and articulable need for the information. If you don’t need the information or know what you plan to do with it, don’t ask for it.

These few questions were geared to the feasibility of extracting data from databases so that it stays utile and complete. Enterprise databases support a raft of standardized reporting capabilities: “screens” or “reports” run to support routine business processes and decision making. An insurance carrier may call a particular report the “Claims File;” but, it is not a discrete “file” at all. It’s a predefined template or report that presents a collection of data extracted from the database in a consistent way. Lots of what we think of as sites or documents are really reports from databases. Your Facebook page? It’s a report. Your e-mail from Microsoft Outlook? Also a report.

In addition to supplying a range of standard reports, enterprise databases can be queried using enhanced reporting capabilities (“custom reports”) and using overlay reporting tools—commercial software “sold separately” and able to interrogate the database in order to produce specialized reporting or support data analytics. A simple example is presentation software that generates handsome charts and graphics based on data in the database. The presentation software didn’t come with the database. It’s something they bought (or built) to “bolt on” for enhanced/overlay reporting.

Although databases are queried using a “query language,” users needn’t dirty their hands with query languages because queries are often executed “under the hood” by the use of those aforementioned standardized screens, reports and templates. Think of these as pre-programmed, pushbutton queries. There is usually more (and often much more) that can be gleaned from a database than what the standardized reports supply, and some of this goes to the integrity of the data itself. In that case, understanding the query language is key to fashioning a query that extracts what you need to know, both within the data and about the data.

As importantly as learning what the database can produce is understanding what the database does or does not display to end users. These are the user experience (UX) and user interface (UI). Screen shots may be worth a thousand words when it comes to understanding what the user saw or what the user might have done to pursue further intelligence.

Enterprise and commercial databases tend to be big and expensive. Accordingly, most are well documented in manuals designed for administrators and end users. When a producing party objects that running a query is burdensome, the manuals may make clear that what you seek is no big deal to obtain.
One feature that sets databases apart from many others forms of ESI is the critical importance of the fielding of data. **Preserving the fielded character of data is essential to preserving its utility and searchability**. “Fielding data” means that information is stored in locations dedicated to holding just that information. Fielding data serves to separate and identify information so you can search, sort and cull using just that information. It’s a capability we take for granted in databases but that is often crippled or eradicated when data is produced in e-discovery. **Be sure that you consider the form of production, and insure that the fielded character of the data produced will not be lost, whether supplied as a standard report or as a delimited export.**

Fielding data isn’t new. We did it back when data was stored as paper documents. Take a typical law firm letter: the letterhead identifies the firm, the date below the letterhead is understood to be the date sent. A *Re:* line follows, denoting matter or subject, then the addressee, salutation, etc. The recipient is understood to be named at the start of the letter and the sender at the bottom. These conventions governing where to place information are vital to our ability to understand and organize conventional correspondence.

Similarly, all of the common productivity file types encountered in e-discovery (Microsoft Office formats, PDF and e-mail) employ fielding to abet utility and functionality. Native “documents” are natively fielded; that is, a file’s content is structured to insure that particular pieces of information reside in defined locations within the file. This structure is understood and exploited by the native application and by tools designed to avail themselves of the file architecture.

We act inconsistently, inefficiently and irrationally when we deal with fielded information in e-discovery. In contrast to just a few years ago, only the most Neanderthal counsel now challenges the need to produce the native fielding of spreadsheet data. Accordingly, production of spreadsheets in native forms has evolved to become routine and (largely) uncontentious. To get to this point, workflows were modified, Bates numbering procedures were tweaked, and despite dire predictions, none of it made the sky fall. We can and must do the same with PowerPoint presentations and Word documents.

“**What’s vice today may be virtue tomorrow,**” wrote novelist (and jurist) Henry Fielding.

Now, take e-mail. All e-mail is natively fielded data, and the architecture of e-mail messages is established by published standards called RFCs—structural conventions that e-mail applications and systems must embrace to insure that messages can traverse any server. The RFCs define placement and labeling of the sender, recipients, subject, date, attachments, routing, message body and other components of every e-mail that transits the Internet.

But when we produce e-mail in discovery, the “accepted” practice is to deconstruct each message and produce it in a cruder fielded format that’s incompatible with the RFCs and unrecognizable to any e-mail tool or system. Too, the production is almost always incomplete compared to the native content.
The deconstruction of fielded data is accomplished by a process called **Field Mapping**. The contents of particular fields within the native source are extracted and inserted into a matrix that may assign the same name to the field as accorded by the native application or rename it to something else altogether. Thus, the source data is “mapped” to a new name and location. At all events, the mapped fields never mirror the field structure of the source file.

*Ever? No, never.*

The jumbled fielding doesn’t entirely destroy the ability to search within fields or cull and sort by fielded content; but, it requires lawyers to rent or buy tools that can re-assemble and read the restructured data in order to search, sort and review the content. And again, information in the original is often omitted, not because it’s privileged or sensitive, but because...well, um, er, *we just do it that way, dammit!*

*But the information that’s omitted, surely that’s useless metadata, right?*

Interestingly, no. In fact, the omitted information significantly aids our ability to make sense of the production, such as the fielded data that allows messages to be organized into conversational threads (e.g., In-Reply-To, References and Message-ID fields) and the fielded data that enables messages to be correctly ordered across time zones and daylight savings time (e.g., UTC offsets).
“Why do producing parties get to recast and omit this useful information,” you ask? The industry responds: "These are not the droids you’re looking for." "Hey, is that Elvis?" "No Sedona for you!"

The real answer is that counsel, and especially requesting counsel, are asleep at the wheel. Producing parties have been getting away with this nonsense, unchallenged, for so long, they’ve come to view it as a birthright. But, reform is coming, at the glacial pace for which we lawyers are justly reviled, I mean revered.

E-discovery standards have indeed evolved to acknowledge that e-mail must be supplied with some fielding preserved; but, there is no sound reason to produce e-mail with shuffled or omitted fields. It doesn't cost more to be faithful to the native or near-native architecture or be complete in supplying fielded content; in fact, producing parties pay more to degrade the production, and what emerges costs more to review.

Perhaps the hardest thing for lawyers and judges to appreciate is the importance fielding plays in culling, sorting and search.

- It’s efficient to be able to cull and sort files only by certain dates.
- It’s efficient to be able to search only within e-mail recipients.
- It’s efficient to be able to distinguish Speaker Notes within a PowerPoint or filter by the Author field in a Word document.

Preserving the fielded character of data makes that possible. Preserving the fielded data and the native file architecture allows use of a broad array of tools against the data, where restructuring fielded data limits its use to only a handful of pricey tools that understand peculiar and proprietary production formats.

It’s not enough for producing parties to respond, “But, you can reassemble the kit of data we produce to make it work somewhat like the original evidence.” In truth, you often can't, and you shouldn’t have to try.

It ties back to the Typewriter Generation mentality that keeps us thinking about “documents” and seeking to define everything we seek as a "document." Most information sought in discovery today is not a purposeful precursor to something that will be printed. Most modern evidence is data, fielded data. Modern productivity files aren’t blobs of text, they're ingenious little databases. Powerful. Rich. Databases. Their native content and architecture are key to their utility and efficient searchability in discovery. Get the fielding right, and functionality follows.

Seeking discovery from databases is a key capability in modern litigation, and it’s not easy for the technically challenged (although it’s probably a whole lot easier than your opponent claims). Getting the proper data in usable forms demands careful thought, tenacity and more-
than-a-little homework. Still, anyone can do it, alone with a modicum of effort, or aided by a little expert assistance.
Search is a Science
The Streetlight Effect in e-Discovery

In the wee hours, a beat cop sees a drunken lawyer crawling around under a streetlight searching for something. The cop asks, “What’s this now?” The lawyer looks up and says, “I’ve lost my keys.” They both search for a while, until the cop asks, “Are you sure you lost them here?” “No, I lost them in the park,” the tipsy lawyer explains, “but the light’s better over here.”

I told that groaner in court, trying to explain why opposing counsel’s insistence that we blindly supply keywords to be run against the e-mail archive of a Fortune 50 insurance company wasn’t a reasonable or cost-effective approach e-discovery. The “Streetlight Effect,” described by David H. Freedman in his 2010 book Wrong, is a species of observational bias where people tend to look for things in the easiest ways. It neatly describes how lawyers approach electronic discovery. We look for responsive ESI only where and how it’s easiest, with little consideration of whether our approaches are calculated to find it.

Easy is wonderful when it works; but looking where it’s easy when failure is assured is something no sober-minded counsel should accept and no sensible judge should allow.

Consider The Myth of the Enterprise Search. Counsel within and without companies and lawyers on both sides of the docket believe that companies have the ability to run keyword searches against their myriad siloes of data: mail systems, archives, local drives, network shares, portable devices, removable media and databases. They imagine that finding responsive ESI hinges on the ability to incant magic keywords like Harry Potter. Documentum Relevantus!

Though data repositories may share common networks, they rarely share common search capabilities or syntax. Repositories that offer keyword search may not support Boolean constructs (queries using “AND,” “OR” and “NOT”), proximity searches (Word1 near Word2), stemming (finding “adjuster,” “adjusting,” “adjusted” and “adjustable”) or fielded searches (restricted to just addressees, subjects, dates or message bodies). Searching databases entails specialized query languages or user privileges. Moreover, different tools extract text and index such extractions in quite different ways, with the upshot being that a document found on one system will not be found on another using the same query.
But the Streetlight Effect is nowhere more insidious than when litigants use keyword searches against archives, e-mail collections and other sources of indexed ESI.

That Fortune 50 company—call it All City Indemnity—collected a gargantuan volume of e-mail messages and attachments in a process called “message journaling.” Journaling copies every message traversing the system into an archive where the messages are indexed for search. Keyword searches only look at the index, not the messages or attachments; so, if you don’t find it in the index, you won’t find it at all.

All City gets sued every day. When a request for production arrives, they run keyword searches against their massive mail archive using a tool we’ll call Truthiness. Hundreds of big companies use Truthiness or software just like it, and blithely expect their systems will find all documents containing the keywords.

They’re wrong...or in denial.

If requesting parties don’t force opponents like All City to face facts, All City and its ilk will keep pretending their tools work better than they do, and requesting parties will keep getting incomplete productions. To force the epiphany, consider the following interrogatory.

**Interrogatory:** For each electronic system or index that will be searched to respond to discovery, please state:

1. The rules employed by the system to tokenize data so as to make it searchable;
2. The stop words used when documents, communications or ESI were added to the system or index;
3. The number and nature of documents or communications in the system or index which are not searchable as a consequence of the system or index being unable to extract their full text or metadata; and
4. Any limitation in the system or index, or in the search syntax to be employed, tending to limit or impair the effectiveness of keyword, Boolean or proximity search in identifying documents or communications that a reasonable person would understand to be responsive to the search.

A court will permit “discovery about discovery” like this when a party demonstrates why an inadequate index is a genuine problem. So, let’s explore the rationale behind each inquiry:

**Tokenization Rules** - When machines search collections of documents for keywords, they rarely search the documents for matches; instead, they consult an index of words extracted from the documents. Machines cannot read, so the characters in the documents are identified as
“words” because their appearance meets certain rules in a process called “tokenization.” Tokenization rules aren’t uniform across systems or software. Many indices simply don’t index short words (e.g., acronyms). None index single letters or numbers.

Tokenization rules also govern such things as the handling of punctuated terms (as in a compound word like “wind-driven”), case (will a search for “roof” also find “Roof?”), diacriticals (will a search for Rene also find René?) and numbers (will a search for “Clause 4.3” work?). Most people simply assume these searches will work. Yet, in many search tools and archives, they don’t work as expected, or don’t work at all, unless steps are taken to ensure that they will work.

**Stop Words** – Some common “stop words” or “noise words” are simply excluded from an index when it’s compiled. Searches for stop words fail because the words never appear in the index. Stop words aren’t always trivial omissions. For example, “all” and “city” were stop words; so, a search for “All City” will fail to turn up documents containing the company’s own name! Words like side, down, part, problem, necessary, general, goods, needing, opening, possible, well, years and state are examples of common stop words. Computer systems typically employ dozens or hundreds of stop words when they compile indices.

Because users aren’t warned that searches containing stop words fail, they mistakenly assume that there are no responsive documents when there may be thousands. A search for “All City” would miss millions of documents at All City Indemnity (though it’s folly to search a company’s files for the company’s name).

**Non-searchable Documents** - A great many documents are not amenable to text search without special handling. Common examples of non-searchable documents are faxes and scans, as well as TIFF images and some Adobe PDF documents. While no system will be flawless in this regard, it’s important to determine how much of a collection isn’t text searchable, what’s not searchable and whether the portions of the collection that aren’t searchable are of particular importance to the case. If All City’s adjusters attached scanned receipts and bids to e-mail messages, the attachments aren’t keyword searchable absent optical character recognition (OCR).

Other documents may be inherently text searchable but not made a part of the index because they’re password protected (i.e., encrypted) or otherwise encoded or compressed in ways that frustrate indexing of their contents. Important documents are often password protected.

**Other Limitations** - If a party or counsel knows that the systems or searches used in e-discovery will fail to perform as expected, they should be obliged to affirmatively disclose such
shortcomings. If a party or counsel is uncertain whether systems or searches work as expected, they should be obliged to find out by, *e.g.*, running tests to be reasonably certain.

No system is perfect, and perfect isn’t the e-discovery standard. Often, we must adapt to the limitations of systems or software. But you have to know what a system can’t do before you can find ways to work around its limitations or set expectations consistent with actual capabilities, not magical thinking and unfounded expectations.
Surefire Steps to Splendid Search

Hear that rumble? It’s the bench’s mounting frustration with the senseless, slipshod way lawyers approach keyword search.

It started with Federal Magistrate Judge John Facciola’s observation that keyword search entails a complicated interplay of sciences beyond a lawyer’s ken. He said lawyers selecting search terms without expert guidance were truly going “where angels fear to tread.”

Federal Magistrate (now District) Judge Paul Grimm called for “careful advance planning by persons qualified to design effective search methodology” and testing search methods for quality assurance. He added that, “the party selecting the methodology must be prepared to explain the rationale for the method chosen to the court, demonstrate that it is appropriate for the task, and show that it was properly implemented.”

Most recently, Federal Magistrate Judge Andrew Peck issued a “wake up call to the Bar,” excoriating counsel for proposing thousands of artless search terms.

Electronic discovery requires cooperation between opposing counsel and transparency in all aspects of preservation and production of ESI. Moreover, where counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI’s custodians as to the words and abbreviations they use, and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of ‘false positives.’ It is time that the Bar—even those lawyers who did not come of age in the computer era—understand this.

No Help
Despite the insights of Facciola, Grimm and Peck, lawyers still don’t know what to do when it comes to effective, defensible keyword search. Attorneys aren’t trained to craft keyword searches of ESI or implement quality control testing for same. And their experience using Westlaw, Lexis or Google serves only to inspire false confidence in search prowess. Even saying “hire an expert” is scant guidance. Who’s an expert in ESI search for your case? A linguistics professor or litigation support vendor? Perhaps the misbegotten offspring of William Safire and Sergey Brin?

The most admired figure in e-discovery search today—the Sultan of Search—is Jason R. Baron at the National Archives and Records Administration, and Jason would be the first to admit he has no training in search. The persons most qualified to design effective search in e-discovery
earned their stripes by spending thousands of hours running searches in real cases—making mistakes, starting over and tweaking the results to balance efficiency and accuracy.

**The Step-by-Step of Smart Search**

So, until the courts connect the dots or better guidance emerges, here’s my step-by-step guide to craftsman like keyword search. I promise these ten steps will help you fashion more effective, efficient and defensible queries.

1. **Start with the Request for Production**
2. **Seek Input from Key Players**
3. **Look at what You’ve Got and the Tools you’ll Use**
4. **Communicate and Collaborate**
5. **Incorporate Misspellings, Variants and Synonyms**
6. **Filter and Deduplicate First**
7. **Test, Test, Test!**
8. **Review the hits**
9. **Tweak the Queries and Retest**
10. **Check the Discards**

**1. Start with the Request for Production**

Your pursuit of ESI should begin at the first anticipation of litigation in support of the obligation to identify and preserve potentially relevant data. Starting on receipt of a request for production (RFP) is starting late. Still, it’s against the backdrop of the RFP that your production efforts will be judged, so the RFP warrants careful analysis to transform its often expansive and bewildering demands to a coherent search protocol.

The structure and wording of most RFPs are relics from a bygone time when information was stored on paper. You’ll first need to hack through the haze, getting beyond the “any and all” and “touching or concerning” legalese. Try to rephrase the demands in everyday English to get closer to the terms most likely to appear in the ESI. Add terms of art from the RFP to your list of keyword candidates. Have several persons do the same, insuring you include multiple interpretations of the requests and obtain keywords from varying points of view.

If a request isn’t clear or is hopelessly overbroad, push back promptly. Request a clarification, move for protection or specially except if your Rules permit same. Don’t assume you can trot out some boilerplate objections and ignore the request. If you can’t make sense of it, or implement it in a reasonable way, tell the other side how you’ll interpret the demand and
approach the search for responsive material. Wherever possible, you want to be able to say, “We told you what we were doing, and you didn’t object.”

2. Seek Input from Key Players

Judge Peck was particularly exercised by the parties’ failure to elicit search assistance from the custodians of the data being searched. Custodians are THE subject matter experts on their own data. Proceeding without their input is foolish. Ask key players, “If you were looking for responsive information, how would you go about searching for it? What terms or names would likely appear in the messages we seek? What kinds of attachments? What distribution lists would have been used? What intervals and events are most significant or triggered discussion?” Invite custodians to show you examples of responsive items, and carefully observe how they go about conducting their search and what they offer. You may see them take steps they neglect to describe or discover a strain of responsive ESI you didn’t know existed.

Emerging empirical evidence underscores the value of key player input. At the latest TREC Legal Track challenge, higher precision and recall seemed to closely correlate with the amount of time devoted to questioning persons who understood the documents and why they were relevant. The need to do so seems obvious, but lawyers routinely dive into search before dipping a toe into the pool of subject matter experts.

3. Look at what You’ve Got and the Tools You’ll Use

Analyze the pertinent documentary and e-mail evidence you have. Unique phrases will turn up threads. Look for words and short phrases that tend to distinguish the communication as being about the topic at issue. What content, context, sender or recipients would prompt you to file the message or attachment in a responsive folder had it occurred in a paper document?

Knowing what you’ve got also means understanding the forms of ESI you must search. Textual content stored in TIFF images or facsimiles demands a different search technique than that used for e-mail container files or word processed documents.

You can’t implement a sound search if you don’t know the capabilities and limitations of your search tool. Don’t rely on what a vendor tells you their tool can do, test it against actual data and evidence. Does it find the responsive data you already know to be there? If not, why not?

Any search tool must be able to handle the most common productivity formats, e.g., .doc, docx, .ppt, .pptx, .xls, .xlsx, and .pdf, thoroughly process the contents of common container files, e.g., .pst, .ost, .zip, and recurse through nested content and e-mail attachments.
As importantly, search tools need to clearly identify any “exceptional” files unable to be searched, such as non-standard file types or encrypted ESI. If you’ve done a good job collecting and preserving ESI, you should have a sense of the file types comprising the ESI under scrutiny. Be sure that you or your service providers analyze the complement of file types and flags any that can’t be searched. Unless you make it clear that certain files types won’t be searched, the natural assumption will be that you thoroughly searched all types of ESI.

4. **Communicate and Collaborate**

Engaging in genuine, good faith collaboration is the most important step you can take to insure successful, defensible search. Cooperation with the other side is not a sign of weakness, and courts expect to see it in e-discovery. Treat cooperation as an opportunity to show competence and readiness, as well as to assess your opponent’s mettle. What do you gain from wasting time and money on searches the other side didn’t seek and can easily discredit? Won’t you benefit from knowing if they have a clear sense of what they seek and how to find it?

Tell the other side the tools and terms you’re considering and seek their input. They may balk or throw out hundreds of absurd suggestions, but there’s a good chance they’ll highlight something you overlooked, and that’s one less do over or ground for sanctions. Don’t position cooperation as a trap nor blindly commit to run all search terms proposed. “We’ll run your terms if you agree to accept our protocol as sufficient” isn’t fair and won’t foster restraint. Instead, ask for targeted suggestions, and test them on representative data. Then, make expedited production of responsive data from the sample to let everyone see what’s working and what’s not.

Importantly, frame your approach to accommodate at least two rounds of keyword search and review, affording the other side a reasonable opportunity to review the first production before proposing additional searches. When an opponent knows they’ll get a second dip at the well, they don’t have to make Draconian demands.

5. **Incorporate Misspellings, Variants and Synonyms**

Did you know Google got its name because its founders couldn’t spell googol? Whether due to typos, transposition, IM-speak, misuse of homophones or ignorance, electronically stored information fairly crawls with misspellings that complicate keyword search. Merely searching for “management” will miss “managment” and “mangement.”

To address this, you must either include common variants and errors in your list of keywords or employ a search tool that supports fuzzy searching. The former tends to be more efficient.
because fuzzy searching (also called *approximate string matching*) mechanically varies letters, often producing an unacceptably high level of false hits.

How do you convert keywords to their most common misspellings and variants? A linguist could help or you can turn to the web. Until a tool emerges that lists common variants and predicts the likelihood of false hits, try a site like [http://www.dumbtionary.com](http://www.dumbtionary.com) that checks keywords against over 10,000 common misspellings and consult Wikipedia’s list of more than 4,000 common misspellings (Wikipedia shortcut: **WP:LCM**).

To identify synonyms, pretend you are playing the board game Taboo. Searches for “car” or “automobile” will miss documents about someone’s “wheels” or “ride.” Consult the thesaurus for likely alternatives for critical keywords, but don’t go hog wild with Dr. Roget’s list. Question key players about internal use of alternate terms, abbreviations or slang.

6. **Filter and Deduplicate First**

Always filter out irrelevant file types and locations before initiating search. Music and images are unlikely to hold responsive text, yet they’ll generate vast numbers of false hits because their content is stored as alphanumeric characters. The same issue arises when search tools fail to decode e-mail attachments before search. Here again, you have to know how your search tool handles encoded, embedded, multibyte and compressed content.

Filtering irrelevant file types can be accomplished various ways, including culling by binary signatures, file extensions, paths, dates or sizes and by de-NISTing for known hash values. The National Institute of Standards and Technology maintains a registry of hash values for commercial software and operating system files that can be used to reliably exclude known, benign files from e-discovery collections prior to search. [http://www.nsrl.nist.gov](http://www.nsrl.nist.gov).

The exponential growth in the volume of ESI doesn’t represent a leap in productivity so much as an explosion in duplication and distribution. Much of the data we encounter are the *same* documents, messages and attachments replicated across multiple backup intervals, devices and custodians. Accordingly, the efficiency of search is greatly aided—and the cost greatly reduced—by *deduplicating* repetitious content *before* indexing data for search or running keywords. Employ a method of deduplication that tracks the origins of suppressed iterations so that repopulation can be accomplished on a per custodian basis.

Applied sparingly and with care, you may even be able to use keywords to *exclude* irrelevant ESI. For example, the presence of keywords “Cialis” or “baby shower” in an e-mail may reliably signal the message isn’t responsive; but *testing and sampling must be used to validate such exclusionary searches*. 

336
7. Test, Test, Test!

The single most important step you can take to assess keywords is to test search terms against representative data from the universe of machines and data under scrutiny. No matter how well you think you know the data or have refined your searches, testing will open your eyes to the unforeseen and likely save a lot of wasted time and money.

The nature and sample size of representative data will vary with each case. The goal in selection isn’t to reflect the average employee’s collection but to fairly mirror the collections of employees likely to hold responsive evidence. Don’t select a custodian in marketing if the key players are in engineering.

Often, the optimum custodial choices will be obvious, especially when their roles made them a nexus for relevant communications. Custodians prone to retention of ESI are better candidates than those priding themselves on empty inboxes. The goal is to flush out problems before deploying searches across broader collections, so opting for uncomplicated samples lessens the value.

It’s amazing how many false hits turn up in application help files and system logs; so early on, I like to test for noisy keywords by running searches against data having nothing whatsoever to do with the case or the parties (e.g., the contents of a new computer). Being able to show a large number of hits in wholly irrelevant collections is compelling justification for limiting or eliminating unsuitable keywords.

Similarly, test search terms against data samples collected from employees or business units having nothing to do with the subject events to determine whether search terms are too generic.

8. Review the Hits

My practice when testing keywords is to generate spreadsheet-style views letting me preview search hits in context, that is, flanked by 20 to 30 words on each side of the hit. It’s efficient and illuminating to scan a column of hits, pinpoint searches gone awry and select particular documents for further scrutiny. Not all search tools support this ability, so check with your service provider to see what options they offer.

Armed with the results of your test runs, determine whether the keywords employed are hitting on a reasonably high incidence of potentially responsive documents. If not, what usages are throwing the search off? What file types are appearing on exceptions lists as unsearchable due to, e.g., obscure encoding, password protection or encryption?
As responsive documents are identified, review them for additional keywords, acronyms and misspellings. Are terms that should be finding known responsive documents failing to achieve hits? Are there any consistent features in the documents with noise hits that would allow them to be excluded by modifying the query?

Effective search is an iterative process, and success depends on new insight from each pass. So expect to spend considerable time assessing the results of your sample search. It’s time wisely invested.

9. Tweak the Queries and Retest

As you review the sample searches, look for ways you can tweak the queries to achieve better precision without adversely affecting recall. Do keyword pairs tend to cluster in responsive documents such that using a Boolean and connector will reduce noise hits? Can you approximate the precise context you seek by controlling for proximity between terms?

If very short (e.g., three letter) acronyms or words are generating too many noise hits, you may improve performance by controlling for case (e.g., all caps) or searching for discrete occurrences (i.e., the term is flanked only by spaces or punctuation).

10. Check the Discards

Keyword search must be judged both by what it finds and what it misses. That’s the “quality assurance” courts demand. A defensible search protocol includes limited examination of the items not generating hits to assess whether relevant documents are being passed over.

Examination of the discards will be more exacting for your representative sample searches as you seek to refine and gain confidence in your queries. Thereafter, random sampling should suffice.

No court has proposed a benchmark or rule-of-thumb for random sampling, but there’s more science to sampling than simply checking every hundredth document. If your budget doesn’t allow for expert statistical advice, and you can’t reach a consensus with the other side, be prepared to articulate why your sampling method was chosen and why it strikes a fair balance between quality assurance and economy. The sampling method you employ needn’t be foolproof, but it must be rational.

Remember that the purpose of sampling the discards is to promptly identify and resolve ineffective searches. If quality assurance examinations reveal that responsive documents are turning up in the discards, those failures must receive prompt attention.
Search Tips

Defensible search strategies are well-documented. Record your efforts in composing, testing and tweaking search terms and the reasons for your choices along the way. Spreadsheets are handy for tracking the evolution of your queries as you add, cut, test and modify them.

Effective searches are tailored to the data under scrutiny. For example, it’s silly to run a custodian’s name or e-mail address against his or her own e-mail, but sensible for other collections. It’s often smart to tier your ESI and employ keywords suited to each tier or, when feasible, to limit searches to just those file types or segments of documents (i.e., message body and subject) likely to be responsive. This requires understanding what you’re searching and how it’s structured.

When searching e-mail for recipients, it’s almost always better to search by e-mail address than by name. In a company with dozens of Bob Browns, each must have a unique e-mail address. Be sure to check whether users employ e-mail aliasing (assigning idiosyncratic “nicknames” to addressees) or distribution lists, as these can thwart search by e-mail address or name.

Search is a Science...

...but one lawyers can master. I guarantee these steps will wring more quality and trim the fat from text retrieval. It's worth the trouble, because the lowest cost e-discovery effort is the one done right from the start.
Exercise 15: Processing, Culling, Search and Export

GOALS: The goals of this exercise are for the student to:

1. Become acquainted with ingestion, processing, culling and search in the context of a commercial e-discovery processing tool; and
2. Generate a Bates-labeled production set with accompanying load files.

OUTLINE: Students will receive a dongle for Nuix, a commercial e-discovery processing tool, and will use Nuix to ingest and process the contents of the forensic image generated in Exercise 3 and the file www.craigball.com/filetypes.zip used in Exercises 7 and 8. Students will then explore the culling, search, analytic and export features of the tool and export an exemplar production set with load files typical of those used in electronic discovery.

Exercise 15a: Ingest and Process Case Evidence

Step 1: Download and install the Nuix software installer suited to your machine (Windows or Mac) from https://download.nuix.com/releases/desktop. Your user name is nuixstudent and your password is Abc123. Be sure to capitalize the “A” in Abc123.

Step 2: Create an empty directory (folder) on your Desktop called My Case and launch Nuix. Acknowledge any missing software notices. Start a new Simple case in your My Case directory:

![Nuix software interface]

66 If the hyperlink doesn’t work, try pasting the URL into the address bar of your browser.
**Step 3:** Add evidence to your case: Select Add>Add Evidence

Supply a name for your evidence group (e.g., “My Evidence”), insure your Source time zone is set correctly and click the “Add” button below the Content box, then select “Add Files:”
From the location where you stored them on your computer, select/open the forensic image (.E01) file you created in Exercise 3 and the file “filetypes.zip (www.craigball.com/filetypes.zip) used in Exercises 7 and 8:

Click “OK” on the Add/Edit Evidence dialogue box and once more on the Add Case Evidence dialogue box. Finally, click “OK” on the Pre-Filter Evidence dialogue box to launch processing. Processing should complete in under five minutes on most machines, resulting in a total of 42,499 processed items. Your completion time may vary, but your item count should be the same (on machines running Windows).

Your “Processing” tab should resemble the screenshot on the next page:

---

67 Your numbers will differ if you used a different thumb drive as your Evidence Drive.
Navigating the Nuix case screen

The Nuix case screen is divided into various tabs and pane that can be resized and separated as desired. Take a moment to locate the following functional regions you will explore:

1. Report tabs
2. Search box
3. Document Navigator
4. Results pane
5. Date Filter
6. Preview pane and Preview pane tabs
7. Export button
8. Review and Tag pane
9. Filtered Items pane
Exercise 15b: Cull and Filter the Collection
Because the cost of e-discovery rises in proportion to the volume of data under review, a unique advantage of e-discovery derives from the ease and speed with which one can cull, filter and de-duplicate a collection. Using Nuix, we will hide immaterial items, deduplicate the remainder and apply date and file type filters.

Immaterial Items: Immaterial items are those extracted for forensic completeness, but having little or no intrinsic value as discoverable evidence. Common examples of immaterial items include the folder structure in which files are stored and the various container files (like mailbox files and compressed file wrappers) that tend to have no relevance apart from their contents.

Step 1: Hide Immaterial Items
Locate the Immaterial Items drop down menu at top of the Results pane and select Hide:

Approximately 4,515 immaterial items are now suppressed.68

Deduplication
The volume of data encountered in e-discovery is largely a function of fragmentation and replication. Fragmentation refers to how, what once might have been addressed in a single business letter of two, may today splay across dozens or hundreds of e-mail messages, attachments and text messages. Replication refers to the ease with which a message or attachment may be dispatched to dozens or hundreds of recipients. Deduplication is a method by which items that are identical in all material respects may be suppressed in favor of a single iteration of same. Only the single instance is then subjected to search and review. Deduplication is typically achieved by comparison of the hash values of files or, in the case of e-mail, by comparing selected constituent parts of messages. Deduplication is termed “vertical” when applied to a single custodian’s collection and “horizontal” or “global” when applied across the collections of multiple custodians.

68 Your numbers will differ if you used a different thumb drive as your Evidence Drive.
Step 2: Deduplicate the Collection
Locate the deduplication drop down menu at top of the Results pane and select “MD5:”

![Results pane showing deduplication drop down menu]

Approximately 12,111 duplicate items are now suppressed.

Step 3: Apply Date and File Type Filters
Locate the date filter drop down menu and select “Before.” Then, enter Jun 1, 2011 as the end date in the right hand date box and click the blue circle with the arrow (below) to apply the temporal filter.

![Date filter applied]

The Results pane should now display approximately 4,851 items.

Locate the Filtered Items pane and, by ticking the box for each, activate filters for Portable Document Format (.pdf), Microsoft Word Document (.doc) and Microsoft Excel Spreadsheet (.xls) (see figure at right).

There should be just 19 instances in the results pane, reflecting a deduplicated collection of .pdf, .doc and .xls files dated before June 1, 2011.

Exercise 15c: Run a Keyword Search
During processing, Nuix extracted all the text it could access in the various files and compiled that text into a searchable database. By running a keyword search, we can further
identify which of the nineteen items in our filtered and deduplicated collection contain the search term “manual.”

Locate the search entry field and type the word “manual” then tap the Enter key on your keyboard:

![Image of search results]

There should now be five items in the Results pane: three .pdf files, one spreadsheet and one Word document. Tick the box at the top of the leftmost column in the Results pane to select these five items.

**Exercise 15d: Generate a Production Set**

If we assume the five filtered items are those that must be produced and that the requesting party sought production of TIFF images accompanied by a Concordance load files, we can now use Nuix to create a Bates-labelled production set to the requesting party’s specification.

**Step 1: Initiate and Configure a Legal Export**

Insure that the check boxes in the leftmost column of the Results pane are checked for each of the fine items to be produced. **Create an empty folder on your Desktop named “Production.”** Click the Export button in the lower right corner of the Results pane and choose “Legal Export to” and “Concordance” from the menus:
The Legal Export menu will appear. Locate the Destination Settings box on the menu under the Export Type tab and set the empty Production folder you’ve just created as your Export Directory. You can use the drop arrow to navigate to the desired destination directory (see illustration on following page)
Select the Numbering and Files tab, and locate the File Naming box. This is where you will choose the forms of production to be exported. If you do not see Native, TIFF and Text already listed in the File Naming box (and you likely won’t), click the “Add…” button, and use the File Type dropdown on the Generated File menu to select Native, TIFF and Text. You will need to do this three times to add native, TIFF and Text as file types. When done correctly, they should appear in the File Naming box, as seen on the next page:

When the Numbering and Files tab of your Legal Export menu looks like the above, click “OK.”
Step 2: Export the Production Set
The next screen you see should be a Pre-Export Summary listing five items for export. If so, click “OK” to proceed to the Exporting Items progress screen and wait to be notified that your five items were successfully exported.

The contents of your Production folder should look like this:
Compare the size of the contents of the TIFF folder to the size of the same items in the Native folder. Enter their sizes below:

Native folder size: __________________    TIFF folder size: __________________
Forms that Function

This article discusses how to request and produce electronically stored information (ESI) in forms that function—that is, in more utile and complete forms of production that preserve the integrity, efficiency and functionality of digital evidence. It explains the advantages of securing production in native and near-native forms, and supplies exemplar language crafted to convey forms of production and metadata values sought.

BACKGROUND

Historically, the law little concerned itself with “forms” of production because there were few alternatives to paper. Then, evidence became digital: documents, pictures, sounds, text messages, e-mail, spreadsheets, presentations, databases and more were created, communicated and recorded as a sequence of “ones” and “zeroes.” Flat forms of information acquired new dimension and depth, described and supplemented by metadata, i.e., data about data supporting the ability to find, use and trust digital information.

Digital photographs hold EXIF data revealing where they were taken and by what camera, spreadsheets carry formulae supporting complex calculations and Word documents store editorial histories and are laced with conversations between collaborators. Presentations feature animated text and rich media, including sound, video and dynamic connections to other data. Databases don’t “store” documents as much as assemble them on demand. Even conversations—once the most ethereal of interactions—now linger as text messages and data packets traversing the internet and cellular networks.

Today, the forms in which information is supplied determine if it is intelligible, functional and complete.

FORMS OF PRODUCTION IN THE FEDERAL RULES

The Federal Rules of Civil Procedure further the goals that lawyers understand the forms of ESI in their cases and resolve forms disputes before requests for production are served. Unresolved forms disputes should be brought to court quickly.

Rule 26(f)(3)(C) requires the parties to submit a discovery plan to the Court prior to the first pretrial conference. The plan must address “any issues about disclosure or discovery of electronically stored information, including the form or forms in which it should be produced.”

Rule 34(b)(1)(C) permits requesting parties to “specify the form or forms in which electronically stored information is to be produced,” yet it’s common for requests for production to be wholly silent on forms of production, despite pages of detailed definitions and instructions.
Practice Tip: Requesting parties should supply a clear and practical written specification of forms sought before the initial Rule 26(f) conference, affording opponents the opportunity to assess the feasibility, cost and burden of producing in specified forms. Even parties who do not know the forms in which an opponent’s data natively resides can anticipate the most common forms of, e.g., e-mail, word processed documents, presentations and spreadsheets.

The Federal Rules lay out **FIVE STEPS** to seeking and objecting to forms of production:

1. Before the first pretrial conference, parties must hash out issues related to “the form or forms in which [ESI] should be produced. FRCP 26(f)(3)(C)

2. Requesting party specifies the form or forms of production for each type of ESI sought: paper, native, near-native, imaged formats or a mix of same. FRCP 34(b)(1)(C)

3. If the responding party will supply the specified forms, the parties proceed with production. If not, the responding party must object and designate the forms in which it intends to make production. If the requesting party fails to specify forms sought, responding party must state the form or forms it intends to produce. FRCP 34(b)(2)(D)

The Notes to Rule 34(b) add: “A party that responds to a discovery request by simply producing electronically stored information in a form of its choice, without identifying that form in advance of the production . . . runs a risk that the requesting party can show that the produced form is not reasonably usable and that it is entitled to production of some or all of the information in an additional form.”

4. If requesting party won’t accept the forms the producing party designates, requesting party must confer with the producing party in an effort to resolve the dispute. FRCP 37(a)(1)

5. If the parties can’t agree, requesting party files a motion to compel, and the Court selects the forms to be produced.

Practice Tip: Even when producing parties use native and near-native forms when reviewing for responsiveness and privilege, the final step before production is often to downgrade the evidence to images before production. Accordingly, requesting parties shouldn’t wait until the response date to ascertain if an opponent refuses to furnish the forms sought. Press for a commitment; and if not forthcoming, move to compel ahead of the response date. Don’t wait to hear the Court ask, “Why didn’t you raise this earlier?”
WHAT ARE THE OPTIONS FOR FORMS OF PRODUCTION?

It’s rarely necessary or advisable to employ a single form of production for all ESI produced in discovery; instead, tailor forms to the data. Options for forms of production include:

- Paper [where the source is paper and the volume small]
- Page Images [best for items requiring redaction and scanned paper records]
- Native [spreadsheets, electronic presentations and word processed documents]
- Near-native [e-mail and database content]
- Hosted production

**Paper**

Converting searchable electronic data to paper is usually a step backward, but paper remains a reasonable choice where the items to be produced are paper documents, few in number and electronic searchability isn’t required.

**Page Images**

Parties produce digital “pictures” of documents, e-mails and other electronic records, typically furnished in Adobe’s Portable Document Format (PDF) or as Tagged Image File Format (TIFF) images. Converting ESI to TIFF images strips its electronic searchability and metadata. Accordingly, TIFF image productions are accompanied by load files holding searchable text and selected metadata. Searchable text is obtained by extraction from an electronic source or for scanned paper documents, by use of optical character recognition (OCR). Load files are composed of delimited text, i.e., values following a predetermined sequence and separated by characters like commas, tabs or quotation marks. The organization of load files must be negotiated, and is often pegged to review software like CT Summation, LexisNexis Concordance or kCura Relativity.

**Pros:** Imaged formats are ideal for production of scanned paper records, microfilm and microfiche, especially when OCR serves to add electronic searchability.

**Cons:** Imaged production breaks down when ESI holds embedded information (e.g., collaborative content like comments or formulae in spreadsheets) or non-printable information (e.g., voice mail, video or animation and structured data). Imaged productions may also serve to degrade evidence when the information is fielded (e.g., structured data and messaging) or functional (e.g., animations in presentations, table relationships in structured data or threads in e-mail).
Native Production
Parties produce the actual data files containing responsive information, e.g., Word documents in their native .DOC or .DOCX formats, Excel spreadsheets as .XLS and .XLSX files and PowerPoint presentations in native .PPT and .PPTX. Native production is cheaper and better in competent hands using tools purpose-built for native review.

Pros: The immediate benefits to the producing party are speed and economy—little or nothing must be spent on image conversion, text extraction or OCR.

The benefits to the requesting party are substantial. Using native review tools or applications like those used to create the data (Careful here!—see Cons below), requesting parties see the evidence as it appeared to the producing party. Embedded commentary and metadata aren’t stripped away, deduplication is facilitated, e-mail messages can be threaded into conversations, time zone irregularities are normalized and costs are reduced and utility enhanced every step of the way.

Cons: Applications needed to view rare and obscure data formats may be prohibitively expensive (e.g., specialized engineering applications or enterprise database software). If native applications are (unwisely) tasked to review, e.g., Microsoft Word for reviewing Word documents, copies must be used to avoid altering evidence.

Near-Native Production
Some ESI cannot be feasibly or prudently tendered in true native formats. Near-native forms preserve the essential utility, content and searchability of native forms but are not, strictly speaking, native forms. Examples:

- **Enterprise e-mail** - When messages are exported from a corporate Exchange mail database to a container format, the container isn’t native to the mail server; but it replicates the pertinent content and essential functionality of the source.
- **Databases** - Exports from databases are often produced in delimited formats not native to the database, yet supporting the ability to interpret the data in ways faithful to the source.
- **Social networking content** - Content from social networking sites like Facebook won’t replicate the precise manner in which the content is stored in the cloud, so near-native forms seek to replicate its essential utility, completeness and searchability.

Hosted Production
Hosted production is more a delivery medium than a discrete form of production. Hosted production resides on a secure website. Requesting parties access data using their web browser, searching, viewing, annotating and downloading data.
MORE ON LOAD FILES
TIFF images cannot carry the text, but PDF images can. Think pants with pockets versus skirts without pockets. When you use TIFF images for production, text has to go somewhere and, since TIFFs have no “pockets,” the text goes into a purse called a “load file.”

Load files first appeared in discovery in the 1980s to add electronic searchability to scanned paper documents and are called load files because they’re used to load data to (“populate”) databases called review platforms.

Different review platforms used different load file formats to order and separate information according to guidelines called “load file specifications.” Load files employ characters called delimiters to field (separate) the various information items in the load file.

Load File Structure
Imagine creating a table to keep track of documents. You might use the first two columns of your table to number the first and last page of each document. The next column holds the document’s name and then each succeeding column carries information about the document. To tell one column from the next, you’d draw lines to delineate the rows and columns, like so:
The lines serve as delimiters—literally delineating one field of data from the next. Vertical and horizontal lines are excellent visual delimiters for humans, but computers work well with characters like commas or tabs. So, if the tabular data were a load file, it might be delimited as:

<table>
<thead>
<tr>
<th>BEGDOC</th>
<th>ENDDOC</th>
<th>FILENAME</th>
<th>MODDATE</th>
<th>AUTHOR</th>
<th>DOCTYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000001</td>
<td>0000004</td>
<td>Contract</td>
<td>01/12/2013</td>
<td>J. Smith</td>
<td>docx</td>
</tr>
<tr>
<td>0000005</td>
<td>0000005</td>
<td>Memo</td>
<td>02/03/2013</td>
<td>R. Jones</td>
<td>docx</td>
</tr>
<tr>
<td>0000006</td>
<td>0000073</td>
<td>Taxes_2013</td>
<td>04/14/2013</td>
<td>H. Block</td>
<td>xlsx</td>
</tr>
<tr>
<td>0000074</td>
<td>0000089</td>
<td>Policy</td>
<td>05/25/2013</td>
<td>A. Dobey</td>
<td>pdf</td>
</tr>
</tbody>
</table>

Each comma replaces a column divider, each line signifies another row and the first or “header” row is used to define the data that follows and the manner in which it’s delimited.

Load files that use commas to separate values are called “comma separated value” or CSV files. More commonly, load files adhere to formats compatible with the Concordance and Summation review tools.
Concordance load files use the file extension .DAT and the þ (thorn, ALT-0254) and ¶ (pilcrow, ALT-0182) characters as delimiters:

**Concordance Load File**

```plaintext
<table>
<thead>
<tr>
<th>BEGDOC</th>
<th>ENDDOC</th>
<th>filename</th>
<th>MODDATE</th>
<th>AUTHOR</th>
<th>DOCTYPE</th>
<th>Page Count</th>
<th>Date Saved</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. Smith</td>
<td>01/12/2013</td>
<td>Contract.docx</td>
<td>0</td>
<td></td>
<td>DOCX</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R. Jones</td>
<td>02/03/2013</td>
<td>Memo.docx</td>
<td>1</td>
<td></td>
<td>DOCX</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Dobey</td>
<td>05/25/2013</td>
<td>Policy.pdf</td>
<td>1</td>
<td></td>
<td>PDF</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Summation load files use the file extension .DII, and separate each record like so:

**Summation Load File**

```plaintext
; Record 1
@T 0000001
@DOCID 0000001
@MEDIA eDoc
@C ENDDOC 0000004
@C PGCOUNT 4
@C AUTHOR J. Smith
@DATESAVED 01/12/2013
@EDOC \NATIVE\Contract.docx
; Record 2
@T 0000005
@DOCID 0000005
@MEDIA eDoc
@C ENDDOC 0000005
@C PGCOUNT 1
@C AUTHOR R. Jones
@DATESAVED 02/03/2013
@EDOC \NATIVE\Memo.docx
@C AUTHOR A. Dobey
@DATESAVED 05/25/2013
@EDOC \NATIVE\Policy.pdf
```

Two more load files:

**Opticon load files** (file extension .OPT) are used in conjunction with Concordance load files to pair Bates numbered pages with corresponding page images and to define the **unitization** of each document; that is, where they begin and end. Document are unitized *physically*, as when
constituent pages are joined by clips, staples or bindings, or *logically*, where constituent pages belong together even if not physically unitized (as when documents are bulk scanned or transmittals reference enclosures). Logical unitization is also a means to track family relationships between container files and contents and e-mail messages and attachments.

Opticon load files employ a simple seven-field, comma-delimited structure:

1. Page identifier,
2. Volume label (optional),
3. Path to page image,
4. New document marker,
5. Box identifier (optional),
6. Folder identifier (optional),
7. Page count (optional).

**Opticon Load File**

```
0000001_0001,TIFF\001\0000001_0001.tif,Y,,4
0000002_0002,TIFF\001\0000002_0002.tif,,,
0000003_0003,TIFF\001\0000003_0003.tif,,,
0000004_0004,TIFF\001\0000004_0004.tif,,,
0000005_0001,TIFF\001\0000005_0001.tif,Y,,1
0000006_0001,TIFF\001\0000006_0001.tif,Y,,68
0000007_0002,TIFF\001\0000007_0002.tif,,,
0000008_0003,TIFF\001\0000008_0003.tif,,,
0000009_0004,TIFF\001\0000009_0004.tif,,,
0000010_0005,TIFF\001\0000010_0005.tif,,,    
```

**Overlay load files** are used to update or correct existing database content by replacing data in fields in the order in which the records occur. Thus, it’s crucial that the order of data within the overlay file match the order of data replaced. Data must be sorted in the same way, and the overlay must not add or omit fields.

**Making the Case against Imaged Production**

Parties don’t print their e-mail before reading it or emboss a document’s name on every page. Parties communicate and collaborate using tracked changes and embedded comments. Parties use native forms because they are the most utile, complete and efficient forms in which to store and access data.

Lawyers come along and convert native forms to images, Bates label each page and purge tracked changes and embedded comments without disclosing the destruction.

Converting a client’s ESI from its native state as kept “in its ordinary course of business” to TIFF images **injects needless expense in at least half a dozen ways:**

1. You pay to convert native forms to TIFF images and emboss Bates numbers;
2. You pay to generate load files;
3. You must produce multiple copies of documents (like spreadsheets) that are virtually incapable of production as images;
4. TIFF images and load files are much “fatter” files than their native counterparts (i.e., bloated 5-40 times as large), so you pay more for vendors to ingest and host them;
5. It’s difficult to reliably de-duplicate documents once converted to images; and
6. You must reproduce everything when opponents recognize that imaged productions fall short of native productions.

REBUTTING THE CASE AGAINST NATIVE

When producing parties insists on converting ESI to TIFF despite a timely request for native production, they often rely on Federal Rules of Civil Procedure 34(b)(2)(E)(ii), which obliges parties to produce ESI in "the form or forms in which it is ordinarily maintained or in a reasonably usable form or forms." This reliance is misplaced because “[i]t is only if the requesting party declines to specify a form that the producing party is offered a choice between producing in the form ‘in which it is ordinary maintained’—native format—or ‘in a reasonably useful form or forms.’ Fed. R. Civ. P. 34(b)(2)(E)(ii)”. The Anderson Living Trust v. WPX Energy Production, LLC, No. CIV 12-0040 JB/LFG. (D. New Mexico March 6, 2014)

Producing parties usually assert FOUR JUSTIFICATIONS for refusing to produce ESI in native and near-native forms. None withstand scrutiny:

1. You can’t Bates label native files. Making the transition to modern forms of production requires acceptance of three propositions:
   • Printouts and images of ESI are not “the same” as ESI;
   • Most items produced in discovery aren’t used in proceedings; and
   • Names of electronic files can be simply changed without altering contents of files.

Native documents carry more information than their imaged counterparts, and are inherently functional, searchable and complete. Moreover, native documents are described by more and different metadata—information invaluable in identifying, sorting and authenticating evidence.

Though you can’t emboss Bates-style identifiers on discrete pages of a native file until printed or imaged, many native forms (e.g., spreadsheets, social networking content, video, and sound files) don’t lend themselves to paged formats and would not be Bates labeled. When Bates-style identifiers are needed on pages for use in proceedings, simply require that file identifiers and page numbers be embossed on images or printouts. In practice, that impacts only a small subset of production.

**Practice tip:** It’s simple and cheap to replace, prepend, or append an incrementing Bates-style identifier to a filename. One free file renaming tool is Bulk Rename Utility, available
2. **Opponents will alter evidence.** Evidence tampering is not a new fear or a hazard unique to e-discovery. Page images, being black and white pictures of text, are simple to manipulate (and Adobe Acrobat has long allowed extensive revision of PDF files).

Though any form of production is prey to unscrupulous opponents, native productions support quick, reliable ways to prevent and detect alteration. Producing native files on read-only media like CDs or DVDs guards against inadvertent alteration. Alterations are easily detected by comparing hash values (digital fingerprints) of suspect files to the files produced.

Counsel savvy enough to seek native production should be savvy enough to refrain from evidence handling practices prone to alter the evidence.

3. **Native production requires broader review.** Native forms routinely hold user-generated content (e.g., collaborative comments in Word documents, animated “off-screen” and layered text in presentations and formulae in spreadsheets) that is rarely visible on page images or intelligible on extracted text. Imaged productions often obliterate such matter *without review and without disclosure, objection or logging*. Review is only “broader” because this user-contributed content has long been furtively and indefensibly stripped away.

4. **Redacting native files changes them.** Change is the sole purpose of redaction. The form of production for items requiring redaction should be the form or forms best suited to efficient removal of privileged or protected content without rendering the remaining content wholly unusable.

Some native file formats support redaction brilliantly; others do not. In the final analysis, the volume of items redacted tends to be insignificant. Accordingly, the form selected for redaction shouldn't dictate the broader forms of production when, overall, native forms have decided advantages for items not requiring.

**Practice Tip:** Don't let the redaction tail wag the production dog. If an opponent wants to redact in .tiff or PDF, let them, but **only** for the redacted items and **only** when they restore searchability after redaction.
UPDATING YOUR REQUESTS FOR PRODUCTION
The first step in getting the information you seek in the forms you desire is to ask for it, applying the rules and eschewing dated boilerplate. Clear, specific requests are the hardest to evade and the easiest to enforce. See Appendix: Exemplar Production Protocol at p. 16, infra.

Most digital evidence—including e-mail—exists as data within databases. So, stop thinking about discovery as the quest for “documents” and start focusing on what you really seek: information in utile and complete forms.

The definition of “document” must give way to an alternate term like “information” or “information items.” Instead of the usual thesaurus-like litany of types of information, consider:

"Information items" as used here encompass individual documents and records (including associated metadata) whether on paper or film, as discrete "files" stored electronically, optically or magnetically or as a record within a database, archive or container file. The term should be read broadly to include e-mail, messaging, word processed documents, digital presentations, spreadsheets and database content.

Next, cut junk prose like “including, but not limited to” and “any and all.” They don’t add clarity. If you must incorporate examples of responsive items in a request, just say “including” and add an instruction that says, “Examples of responsive items set out in any request should not be construed to limit the scope of the request.” If drafting a request without “any and all” makes you quake, add the instruction, “Requests for production should be read so as to encompass any and all items responsive to the request.”

Before you serve discovery, check your definitions to be sure you’ve defined only terms you’ve used and used terms only in ways consistent with your definitions.

Specify the forms you seek
The most common error seen in requests for production is the failure to specify the forms sought for ESI production. Worse, requests often contain legacy boilerplate specifying forms the requesting party doesn’t want.

Every request for production should specify forms of production sensibly and precisely. Don’t assume that “native format” is clear or sufficient; instead, specify the formats sought for common file types, e.g.:
Information that exists in electronic form should be produced in native or near-native formats and should not be converted to imaged formats. Native format requires production in the same format in which the information was customarily created, used and stored in the ordinary course. The table below supplies examples of the native or near-native forms in which specific types of electronically stored information (ESI) should be produced.

<table>
<thead>
<tr>
<th>Source ESI</th>
<th>Native or Near-Native Form or Forms Sought</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Word documents</td>
<td>.DOC, .DOCX</td>
</tr>
<tr>
<td>Microsoft Excel Spreadsheets</td>
<td>.XLS, .XLSX</td>
</tr>
<tr>
<td>Microsoft PowerPoint Presentations</td>
<td>.PPT, .PPTX</td>
</tr>
<tr>
<td>Microsoft Access Databases</td>
<td>.MDB, .ACCDB</td>
</tr>
<tr>
<td>WordPerfect documents</td>
<td>.WPD</td>
</tr>
<tr>
<td>Adobe Acrobat Documents</td>
<td>.PDF</td>
</tr>
<tr>
<td>Images</td>
<td>.JPG, JPEG, .PNG</td>
</tr>
<tr>
<td>E-mail</td>
<td>Messages should be produced in a form or forms that readily support import into standard e-mail client programs; that is, the form of production should adhere to the conventions set out in the internet e-mail standard, RFC 5322. For Microsoft Exchange or Outlook messaging, .PST format will suffice. Single message production formats like .MSG or .EML may be furnished with folder data. For Lotus Notes mail, furnish .NSF files or convert to .PST. If your workflow requires that attachments be extracted and produced separately from transmitting messages, attachments should be produced in their native forms with parent/child relationships to the message and container(s) preserved and produced in a delimited text file.</td>
</tr>
<tr>
<td>Databases</td>
<td>Unless the entire contents of a database are responsive, extract responsive content to a fielded and electronically searchable format preserving metadata values, keys and field relationships. If doing so is infeasible, please identify the database and supply information concerning the schema and query language of the database along with a detailed description of its export capabilities so as to facilitate crafting a query to extract and export responsive data.</td>
</tr>
</tbody>
</table>

362
Documents that do not exist in native electronic formats or which require redaction of privileged content should be produced in searchable .PDF formats or as single page .TIFF images with unredacted OCR text furnished and logical unitization and family relationships preserved.

**Practice Tip:** In settling upon a form of production for e-mail, use this inquiry as a litmus test to distinguish “native” forms from less functional forms: **Can the form produced be imported into common e-mail client or server applications?** If the form of the e-mail is so degraded that e-mail programs cannot recognize it as e-mail, that’s a strong indication the form of production has strayed too far from functional.

**Specify the Load File Format**

Every electronic file has a complement of descriptive information called *system metadata* residing in the file table of the system or device storing the file. Different file types have different metadata. Every e-mail message has “fields” of information in the message “header” that support better searching, sorting and organization of messages. This may be data probative in its own right or simply advantageous to managing and authenticating electronic evidence. Either way, you want to be certain to request it sensibly and precisely. Simply demanding “the metadata” reveals you don’t fully understand what you’re seeking.

Develop a comprehensive production protocol tailored to the case and serve same with discovery. Always specifically request the metadata and header fields you seek, *e.g.*:

Produce delimited load file(s) supplying relevant system metadata field values for each information item by Bates number. Typical field values supplied include:

a. **Source file name** (original name of the item or file when collected from the source custodian or system);
b. **Source file path** (fully qualified file path from the root of the location from which the item was collected);
c. **Last modified date and time** (last modified date and time of the item);
d. **UTC Offset** (The UTC/GMT offset of the item’s modified date and time, *e.g.*, -500).
e. **Custodian or source** (unique identifier for the original custodian or source);
f. **Document type**;
g. **Production File Path** (file path to the item from the root of the production media);
h. **MD5 hash** (MD5 hash value of the item as produced);
i. **Redacted flag** (indication whether the content or metadata of the item has been altered after its collection from the source custodian or system);
j. **Embedded Content Flag** (indication that the item contains embedded or hidden comments, content or tracked changes); and
k. **Deduplicated instances** (by full path).

The following additional fields shall accompany production of e-mail messages:

363
l. To (e-mail address(es) of intended recipient(s) of the message);
m. From (e-mail address of the person sending the message);
n. CC (e-mail address(es) of person(s) copied on the message);
o. BCC (e-mail address(es) of person(s)blind copied on the message);
p. Subject (subject line of the message);
q. Date Received (date the message was received);
r. Time Received (time the message was received);
s. Attachments (beginning Bates numbers of attachments);
t. Mail Folder Path (path of the message from the root of the mail folder); and
u. Message ID (unique message identifier).

Hybrid productions mixing mix imaged and native formats also require that paths to images and extracted text be furnished, as well as logical unitization data serving as the electronic equivalent of paper clips and staples.

De-duplication and Redaction
You may wish to specify whether the production should or should not be de-duplicated, e.g.:

Documents should be vertically de-duplicated by custodian using each document’s hash value. Near-deduplication should not be employed so as to suppress different versions of a document, notations, comments, tracked changes or application metadata.

Because redaction tends to impact just a small part of most productions, it’s important that it not co-opt the forms of production.

Information items that require redaction shall be produced in static image formats, e.g., single page .tiff or multipage PDF images with logical unitization preserved. The unredacted content of each document should be extracted by optical character recognition (OCR) or other suitable method to a searchable text file produced with the corresponding page image(s) or embedded within the image file. Redactions should not be accomplished in a manner that serves to downgrade the ability to electronically search the unredacted portions of the item.

A TIFF-OCR redaction method works reasonably well for text documents, but often fails when applied to complex and dynamic documents like spreadsheets and databases. Unlike text, you can’t spellcheck numbers, so the inevitable errors introduced by OCR make it impossible to have confidence in numeric content or reliably search the data. Moreover, converting a spreadsheet to a TIFF image strips away its essential functionality by jettisoning the underlying formulae that distinguishes a spreadsheet from a table.
Specify the medium of production
A well-crafted request should address the medium of ESI production; that is the mechanism used to convey the electronic production to the requesting party. If you’re receiving 100GB of data, you don’t want it tendered on 143 CDs.

Production of ESI should be made using appropriate electronic media of the producing party’s choosing that does not impose an undue burden or expense upon a recipient.

Conclusion
It’s time to take a hard look at the language of the definitions and instructions accompanying requests for production. Most are boilerplate borrowed from someone who borrowed it from someone who drafted it in 1947. It’s hand-me-down verbiage long past retirement age; so, retire it and craft modern requests for a modern digital world.

We will never be less digital than we are today. Isn’t it time we demand modern evidence and obtain it in the forms in which it serves us best? We must move forms of production upstream, from depleted images and load files to functional native and near native forms retaining the content and structure that supports migration into any form. Utile forms. Complete forms. Forms that function.
Exemplar Production Protocol

This Appendix is an example of a production protocol, sometimes called a data delivery standard. Geared to civil litigation and seeking the lowest cost approach to production of ESI, it seeks native production of common file types and relieves parties of the burden to convert ESI to imaged formats except when needed for redaction. This exemplar protocol specifies near-native alternatives for production of native forms when near-native forms are preferable. For an example of a U.S. Government data delivery standard, see: http://www.sec.gov/divisions/enforce/datadeliverystandards.pdf

Appendix: Exemplar Production Protocol

1. "Information items" as used here encompass individual documents and records (including associated metadata) whether on paper or film, as discrete "files" stored electronically, optically or magnetically or as a record within a database, archive or container file. The term should be read broadly to include e-mail, messaging, word processed documents, digital presentations, spreadsheets and database content.

2. Information that exists in electronic form should be produced in native formats and should not be converted to imaged formats. Native format requires production in the same format in which the information was customarily created, used and stored in the ordinary course.

3. If it is infeasible to produce an item of responsive ESI in its native form, it may be produced in an agreed-upon near-native form; that is, in a form in which the item can be imported into the native application without a material loss of content, structure or functionality as compared to the native form. Static image production formats serve as near-native alternatives only for information items that are natively static images (i.e., photographs and scans of hard-copy documents).

4. The table below supplies examples of agreed-upon native or near-native forms in which specific types of ESI should be produced:

<table>
<thead>
<tr>
<th>Source ESI</th>
<th>Native or Near-Native Form or Forms Sought</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Word documents</td>
<td>.DOC, .DOCX</td>
</tr>
<tr>
<td>Microsoft Excel Spreadsheets</td>
<td>.XLS, .XLSX</td>
</tr>
<tr>
<td>Microsoft PowerPoint</td>
<td>.PPT, .PPTX</td>
</tr>
<tr>
<td>Presentations</td>
<td></td>
</tr>
<tr>
<td>Microsoft Access Databases</td>
<td>.MDB, .ACCDB</td>
</tr>
<tr>
<td>WordPerfect documents</td>
<td>.WPD</td>
</tr>
<tr>
<td>Adobe Acrobat Documents</td>
<td>.PDF</td>
</tr>
<tr>
<td>Photographs</td>
<td>.JPG, .PDF</td>
</tr>
<tr>
<td>E-mail</td>
<td>Messages should be produced in a form or forms that readily support import into standard e-mail</td>
</tr>
</tbody>
</table>
client programs; that is, the form of production should adhere to the conventions set out in the internet e-mail standard, RFC 5322. For Microsoft Exchange or Outlook messaging, .PST format will suffice. Single message production formats like .MSG or .EML may be furnished with folder data. For Lotus Notes mail, furnish .NSF files or convert to .PST. If your workflow requires that attachments be extracted and produced separately from transmitting messages, attachments should be produced in their native forms with parent/child relationships to the message and container(s) preserved and produced in a delimited text file.

<table>
<thead>
<tr>
<th>Databases</th>
<th>Unless the entire contents of a database are responsive, extract responsive content to a fielded and electronically searchable format preserving metadata values, keys and field relationships. If doing so is infeasible, please identify the database and supply information concerning the schema and query language of the database along with a detailed description of its export capabilities so as to facilitate crafting a query to extract and export responsive data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents that do not exist in native electronic formats or which require redaction of privileged content should be produced in searchable .PDF formats or as single page .TIFF images with OCR text of unredacted content furnished and logical unitization and family relationships preserved.</td>
<td></td>
</tr>
</tbody>
</table>

5. Absent a showing of need, a party shall produce responsive information reports contained in databases through the use of standard reports; that is, reports that can be generated in the ordinary course of business and without specialized programming efforts beyond those necessary to generate standard reports. All such reports shall be produced in a delimited electronic format preserving field and record structures and names. The parties will meet and confer regarding programmatic database productions as necessary.

6. Information items that are paper documents or that require redaction shall be produced in static image formats scanned at 300 dpi e.g., single-page Group IV.TIFF or multipage PDF images. If an information item employs color to convey information (versus purely decorative use), the producing party shall not produce the item in a form that does not display color. The full content of each document will be extracted directly from the native source where
feasible or, where infeasible, by optical character recognition (OCR) or other suitable method
to a searchable text file produced with the corresponding page image(s) or embedded within
the image file. Redactions shall be logged along with other information items withheld on
claims of privilege.

7. Parties shall take reasonable steps to ensure that text extraction methods produce usable,
accurate and complete searchable text.

8. Individual information items requiring redaction shall (as feasible) be redacted natively,
produced in .PDF format and redacted using the Adobe Acrobat redaction feature or redacted
and produced in another reasonable manner that does not serve to downgrade the ability to
electronically search the unredacted portions of the item. Bates identifiers should be
endorsed on the lower right corner of all images of redacted items so as not to obscure
content.

9. Upon a showing of need, a producing party shall make a reasonable effort to locate and
produce the native counterpart(s) of any .PDF or .TIF document produced. The parties agree
to meet and confer regarding production of any such documents. This provision shall not
serve to require a producing party to reveal redacted content.

10. Except as set out in this Protocol, a party need not produce identical information items in
more than one form. The content, metadata and utility of an information item shall all be
considered in determining whether information items are identical, and items reflecting
different information shall not be deemed identical.

11. Production of ESI should be made using appropriate electronic media of the producing party’s
choosing that does not impose an undue burden or expense upon a recipient. Label all media
with the case number, production date, Bates range and disk number (1 of X, if applicable).
Organize productions by custodian, unless otherwise instructed. All productions should be
encrypted for transmission to the receiving party. The producing party shall,
contemporaneously with production, separately supply decryption credentials and
passwords to the receiving party for all items produced in an encrypted or password-
protected form.

12. Each information item produced shall be identified by naming the item to correspond to a
Bates identifier according to the following protocol:

i. The first four (4) characters of the filename will reflect a unique alphanumeric
designation identifying the party making production;
ii. The next six (6) characters will be a designation reserved to the discretionary use of the party making production for the purpose of, e.g., denoting the case or matter. This value shall be padded with leading zeroes as needed to preserve its length;

iii. The next nine (9) characters will be a unique, consecutive numeric value assigned to the item by the producing party. This value shall be padded with leading zeroes as needed to preserve its length;

iv. The final six (6) characters are reserved to a sequence consistently beginning with a dash (-) or underscore (_) followed by a five digit number reflecting pagination of the item when printed to paper or converted to an image format for use in proceedings or when attached as exhibits to pleadings.

v. By way of example, a Microsoft Word document produced by Acme in its native format might be named: ACMESAMPLE000000123.docx. Were the document printed out for use in deposition, page six of the printed item must be embossed with the unique identifier ACMESAMPLE000000123_00006. Bates identifiers should be endorsed on the lower right corner of all printed pages, but not so as to obscure content.

vi. This format of the Bates identifier must remain consistent across all productions. The number of digits in the numeric portion and characters in the alphanumeric portion of the identifier should not change in subsequent productions, nor should spaces, hyphens, or other separators be added or deleted except as set out above.

13. Information items designated Confidential may, at the Producing Party’s option:

   a. Be separately produced on electronic production media prominently labeled to comply with the requirements of the [DATE] Protective Order entered in this matter; or, alternatively,

   b. Each such designated information item shall have appended to the file’s name (immediately following its Bates identifier) the following protective legend: `CONFIDENTIAL-SUBJ_TO_PROTECTIVE_ORDER

When any item so designated is converted to a printed or imaged format for use in any submission or proceeding, the printout or page image shall bear the protective legend on each page in a clear and conspicuous manner, but not so as to obscure content.

14. Producing party shall furnish a delimited load file supplying the metadata field values listed below for each information item produced (to the extent the values exist and as applicable):
<table>
<thead>
<tr>
<th>Field Name</th>
<th>Sample Data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BegBates</td>
<td>ACMESAMPLE0000000001</td>
<td>First Bates identifier of item</td>
</tr>
<tr>
<td>EndBates</td>
<td>ACMESAMPLE0000000123</td>
<td>Last Bates identifier of item</td>
</tr>
<tr>
<td>AttRange</td>
<td>ACMESAMPLE0000000124 - ACMESAMPLE000000130</td>
<td>Bates identifier of the first page of the parent document to the Bates identifier of the last page of the last attachment “child” document</td>
</tr>
<tr>
<td>BegAttach</td>
<td>ACMESAMPLE0000000124</td>
<td>First Bates identifier of attachment range</td>
</tr>
<tr>
<td>EndAttach</td>
<td>ACMESAMPLE000000130</td>
<td>Last Bates identifier of attachment range</td>
</tr>
<tr>
<td>Parent_Bates</td>
<td>ACMESAMPLE0000000001</td>
<td>First Bates identifier of parent document/e-mail message. **This Parent_Bates field should be populated in each record representing an attachment “child” document. **</td>
</tr>
<tr>
<td>Child_Bates</td>
<td>ACMESAMPLE0000000004; ACMESAMPLE000000012; ACMESAMPLE000000027</td>
<td>First Bates identifier of “child” attachment(s); may be more than one Bates number listed depending on number of attachments. **The Child_Bates field should be populated in each record representing a “parent” document. **</td>
</tr>
<tr>
<td>Custodian</td>
<td>Houston, Sam</td>
<td>E-mail: mailbox where the email resided. Native: Individual from whom the document originated</td>
</tr>
<tr>
<td>Path</td>
<td>E-mail: \Deleted Items\Battles\SanJac.msg Native: Z:\TravisWB\Alamo.docx</td>
<td>E-mail: Original location of e-mail including original file name. Native: Path where native file document was stored including original file name.</td>
</tr>
<tr>
<td>From</td>
<td>E-Mail: <a href="mailto:Davy@Crockett.net">Davy@Crockett.net</a> Native: D. Crockett</td>
<td>E-mail: Sender Native: Author(s) of document **semi-colons separate multiple entries **</td>
</tr>
<tr>
<td>To</td>
<td>Gen. A.L. de Santa Anna [mailto: <a href="mailto:sa@sa.mx">sa@sa.mx</a>]</td>
<td>Recipient(s) **semi-colons separate multiple entries **</td>
</tr>
<tr>
<td>CC</td>
<td><a href="mailto:Jim.Bowie@bigknife.com">Jim.Bowie@bigknife.com</a></td>
<td>Carbon copy recipient(s) **semi-colons separate multiple entries **</td>
</tr>
<tr>
<td>BCC</td>
<td><a href="mailto:AustinSF@state.tx.gov">AustinSF@state.tx.gov</a></td>
<td>Blind carbon copy recipient(s) **semi-colons separate multiple entries **</td>
</tr>
<tr>
<td>Date Sent</td>
<td>03/18/2015</td>
<td>E-mail: Date the email was sent</td>
</tr>
<tr>
<td>Time Sent</td>
<td>11:45 AM</td>
<td>E-mail: Time the message was sent</td>
</tr>
<tr>
<td>Subject/Title</td>
<td>Remember the Alamo!</td>
<td>E-mail: Subject line of the message</td>
</tr>
<tr>
<td>IntMsgID</td>
<td><a href="mailto:A1315BC17ABD4774BF779CB3E3E62B9B@gmail.com">A1315BC17ABD4774BF779CB3E3E62B9B@gmail.com</a></td>
<td>E-mail: For e-mail in Microsoft Outlook/Exchange, the “Unique Message ID” field; For e-mail in Lotus Notes, the UNID field. Native: empty.</td>
</tr>
<tr>
<td>Date_Mod</td>
<td>02/23/2015</td>
<td>E-mail: empty. Native: Last Modified Date</td>
</tr>
<tr>
<td>Time_Mod</td>
<td>01:42 PM</td>
<td>E-mail: empty Native: Last Modified Time</td>
</tr>
<tr>
<td>File_Type</td>
<td>XLSX</td>
<td>E-mail: empty Native: file type</td>
</tr>
<tr>
<td>Redacted</td>
<td>Y</td>
<td>Denotes that item has been redacted as containing privileged content (yes/no).</td>
</tr>
<tr>
<td>File_Size</td>
<td>1,836</td>
<td>Size of native file document/email in KB.</td>
</tr>
<tr>
<td>HiddenCnt</td>
<td>N</td>
<td>Denotes presence of hidden Content/Embedded Objects in item(s) (Y/N)</td>
</tr>
<tr>
<td>-----------</td>
<td>---</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Confidential</td>
<td>Y</td>
<td>Denotes that item has been designated as confidential pursuant to protective order (Y/N).</td>
</tr>
<tr>
<td>MD5_Hash</td>
<td>eb71a966dcdddb929c1055ff2f1ccd5b</td>
<td>MDS Hash value of the item.</td>
</tr>
<tr>
<td>DeDuped</td>
<td>E-mail: \Inbox\SanJac.msg Native: Z:\CrockettD\Alamo.docx</td>
<td>Full path of deduped instances. <strong>semi-colons separate multiple entries</strong></td>
</tr>
</tbody>
</table>

15. Each production should include a cross-reference load file that correlates the various files, images, metadata field values and searchable text produced.

16. Parties shall respond to each request for production by listing the Bates identifiers/ranges of responsive documents produced, and where an information item responsive to these discovery requests has been withheld or redacted on a claim that it is privileged, the producing party shall furnish a privilege log.
Exercise 16: Forms of Production: Load Files

GOALS: The goals of this exercise are for the student to explore the purpose and structure of load files as used to transmit metadata and communicate the organization of production deliverables to e-discovery review platforms.

OUTLINE: Students will download a compressed file holding four different delimited load file formats (comma separated values-CSV, tab separated values-TXT, Concordance-DAT and Summation-DII), each representing an identical production of e-mail and loose documents. Students will compare the plain text content of the four files in a text viewer then import each into Excel, using the delimiters to parse the load file data into properly populated rows and labeled columns.

Note: if you don’t have a copy of Excel or another spreadsheet program, you can complete the parts of the exercise examining CSV and TSV formats using the free spreadsheet application at Google Drive (http://drive.google.com).

Exercise 16A (15 minutes):

Step 1: Decompress the Files

Download the Zip file at www.craigball.com/GT_loadfile.zip and extract its contents to your desktop or any other convenient location on your computer. Locate the following four files:

1. GT_loadfile.csv (comma separated values);
2. GT_loadfile.txt (tab separated values);
3. GT_loadfile.dat (Concordance format); and
4. GT_loadfile.dii (Summation format).

Step 2: View the Contents of Each File

Run any simple text editor or viewer application and open each file to examine its contents.

In Windows: Use the Wordpad application.
In MacOS: Use the Mac TextEdit application.

Note that each of these files conveys the same tabular information about the items in the production. The differences between them stem from differences in their structure and their use of different delimiters (i.e., field separators).
If successful, you should see content for each like that in the following screen shots:

**GT_loadfile.csv**

![Image of GT_loadfile.csv]

**GT_loadfile.txt**

![Image of GT_loadfile.txt]

**GT_loadfile.dat**

![Image of GT_loadfile.dat]

**GT_loadfile.dii**

![Image of GT_loadfile.dii]
**Question 1:** How many total items (files, attachments and messages) were produced with these load files? **HINT:** the highest value DOCID corresponds with the total number of discrete items produced.

Answer: ________________

**Question 2:** What delimiter(s) are used to separate data fields in each record of the load file?

Answers:

a. GT_loadfile.csv: ________________________

b. GT_loadfile.txt: ________________________

c. GT_loadfile.dat (Concordance format): ________________________

d. GT_loadfile.dii (Summation format): ________________________

**Question 3:** In a text editor application (e.g., Notepad or Word), what character appears when you type **0254** on the numeric keypad of your keyboard while depressing the **ALT** key? What about when you type **ALT-0182**?

ALT-0254 Answer: ________________________

ALT-0182 Answer: ________________________

**Question 4:** These characters have unique names. What are they?

ALT-0254 Answer: ________________________

ALT-0182 Answer: ________________________

*Hint: Neither answer is “the paragraph sign.” Don’t spend more than five minutes on Question 4. GIYF!*
Run Microsoft Excel or your preferred spreadsheet application). Click on the “Data” tab on the menu bar and select “From Text.” The “Import Text File” menu should appear. Locate the “File Name” box and, immediately to its right, click the drop down menu and choose the option, “All Files (*.*)” as below.

Now, browse to the location where you stored the load files you unzipped and select the file named “GT_loadfile.csv,” then click “Open.”

If the Text Import Wizard appears, choose “Delimited” as the file type that best describe your data (see figure below) and click “Next.”
In the next screen of the Wizard, check “Comma,” and be sure no other delimiters are checked. (see figure below). The Data preview window allows you to see how the data will be divided as columns and rows. Click “Finish.”

Click “OK” on the Import Data menu, as below:
Step 2: Convert the TXT-Formatted Load File Content to Tabular Data

Open a new blank workbook in Excel (File>New>Blank Workbook). Follow the same steps as above except open the file called “GT_loadfile.txt” and this time select Tab as your preferred delimiter, and be sure no other delimiters are checked.

Step 3: Convert the Concordance-Formatted Load File Content to Tabular Data

Open a new blank workbook in Excel (File>New>Blank Workbook). Follow the same steps as above except open the file called “GT_loadfile.dat” and this time select Other as your preferred delimiter and, after clicking in the small white box alongside “Other,” enter the ALT-0254 character, and be sure no other delimiters are checked.

Your screen should look like the following before proceeding. If it does, selecting Next and Finish, as before. Save each of the three Excel spreadsheets you’ve created.

Discussion Question: Consider the consequences of a misplaced or omitted delimiter in terms of the reliability of a load file?
Exercise 17: Forms of Production and Cost

GOALS: The goals of this exercise are for the student to:

1. Convert evidence to PDF and TIFF with text; and
2. Assess impact of alternate forms of production in terms of impact on cost of ingestion and hosting.

OUTLINE: Students will convert a Microsoft Word document to PDF, TIFF and text formats, compare file sizes and calculate the projected cost of ingestion and monthly hosting for alternate forms of production when the cost of services is assessed on a per-gigabyte pricing model.

Producing parties frequently seek to convert native file formats used by and collected from custodian into static image formats like PDF or more commonly, TIFF images plus load files holding extracted text or text generated through use of optical character recognition. Proponents of static image productions assert claims of superior document security and point to the ability to emboss page numbers and other identifiers on page images. Too, page images can be viewed using any browser application, affording users ready accessibility to some content, albeit sacrificing other content and utility.

Often overlooked in the debate over forms of production is the impact on ingestion, processing, storage and export costs engendered by use of static image formats. Most e-discovery service providers charge to ingest, process, host (store) and export electronically stored information on a per-gigabyte basis. As a result, when items produced occupy more space (measured in bytes), they cost the recipient more to use. This exercise invites students to consider what, if any, increase in cost may flow from the use of static imaged formats as forms of production.

The Myth of Page Equivalency

It's comforting to quantify electronically stored information as some number of pieces of paper or bankers' boxes. Paper and lawyers are old friends. But you can't reliably equate a volume of data with a number of pages unless you know the composition of the data. Even then, it's a leap of faith.

If you troll the Internet for page equivalency claims, you'll be astounded by how widely they vary, though each is offered with utter certitude. A gigabyte of data is variously equated to an absurd 500 million typewritten pages, a naively accepted 500,000 pages, the popularly cited 75,000 pages and a laggardly 15,000 pages. The other striking aspect of page equivalency claims is that they're
blithely accepted by lawyers and judges who wouldn't concede the sky is blue without a supporting string citation.

In testimony before the committee drafting the federal e-discovery rules, Exxon Mobil representatives twice asserted that one gigabyte yields 500,000 typewritten pages. The National Conference of Commissioners on Uniform State Laws proposes to include that value in its "Uniform Rules Relating to Discovery of Electronically Stored Information." The Conference of Chief Justices cites the same equivalency in its "Guidelines for State Trial Courts Regarding Discovery of Electronically-stored Information." Scholarly articles and reported decisions pass around the 500,000 pages per gigabyte value like a bad cold. Yet, 500,000 pages per gigabyte isn't right. It's not even particularly close to right.

Years ago, Kenneth Withers, Deputy Executive Director of The Sedona Conference and then e-discovery guru for the Federal Judicial Center, wrote a section of the fourth edition of "The Manual on Complex Litigation" that equated a terabyte of data to 500 billion typewritten pages. It was supposed to say million, not billion. Eventually, the typo was noticed and corrected; but, the echoes of that innocent thousand-fold mistake still reverberate today. Anointed by the prestige of the manual, the 500-billion-page equivalency was embraced as gospel. Even when the value was "corrected" to 500 million pages per terabyte—equal to 500,000 pages per gigabyte—we're still talking about equivalency with all the credibility of an Elvis sighting.

So, how many pages are there in a gigabyte? It's the answer lawyers love: “It depends.”

Page equivalency is a myth. One must always look at individual file types and quantities to gauge page equivalency, and there is no reliable rule of thumb geared to how many files of each type a typical user stores. It varies by industry, by user and even by the life span of the media and the evolution of particular applications. A reliable page equivalency must be expressed with reference to both the quantity and form of the data, e.g., "a gigabyte of single page TIF images of 8-1/2-inch x 11-inch documents scanned at 300 dots per inch equals approximately 18,000 pages."

Exercise 17a: Convert Word Document to Imaged Formats

For this exercise, you will download an exemplar Word document and use free, online tools to convert the file to PDF, TIFF and plain text formats.

Step 1: Download the File. Download the file http://www.craigball.com/Always_and_Never.docx and save it to your Desktop or some other location where you can easily find it for this exercise. Should your system not permit download of Word files, you can download the file as a compressed .Zip file from here. Be sure to extract the .DOCX form of the file to your Desktop before proceeding. You must undertake the conversion exercise using the .DOCX form of the file.
Step 2. Convert the .DOCX file to a PDF. Though there are many ways to convert a Word document to a PDF format, including by using Word itself to Save As a PDF or Print to PDF, we will use an online file converter here for consistency and simplicity.

Using your browser, go to https://convertio.co/convert-docx/ and click on the red SELECT YOUR FILES button.

From the Select Files to Convert screen, select “Choose from Computer” then navigate to the file just downloaded called Always_and_Never.docx. Select the file and click “Open.”

You should see the following screen:

Note the pulldown menu where you may select the format for conversion (JPG in the figure at right) and select the down arrow to view options.

Select DOCUMENT and PDF from the menu and submenu (see figure at right).

Click the red Convert button.

In the next screen, click the green DOWNLOAD button and save the Always_and_Never.PDF file to the same location where you saved the .DOCX file.

Step 3: Convert the .DOCX file to TIFF images. Follow the same steps as above, but this time select IMAGE>TIFF using the drop down menu (see image below) before clicking the red “CONVERT” button.
Click the green DOWNLOAD button again and save the file Always_and_Never.tiff to the same location where you placed the .DOCX and .PDF files.

**Step 4: Convert the .DOCX file to plain text.** Follow the same steps, but now select DOCUMENT>TXT from the drop down menu (see image below) before clicking the red “CONVERT” button.

Click the green DOWNLOAD button again and save the file Always_and_Never.txt to the same location where you placed the .DOCX and .PDF files.

**Step 5: Record the file sizes.** Navigate to the location where you downloaded the files and record their file sizes in the blanks below. *Be sure to note if the size value is expressed in units of bytes, kilobytes, megabytes or gigabytes.*

Always_and_Never.DOCX: 18.3 KB

Always_and_Never.PDF: 627 KB

Always_and_Never.TIFF: ______ MB

Always_and_Never.TXT: ______ KB

**Exercise 17b: Calculate the Cost Difference Flowing from Alternate Forms of Production**

There may be many variables that go into computing the cost of vendor services for e-discovery, and the charges for ingestion, processing, hosting and export are just parts of a more complicated puzzle. The purpose of this exercise is to gauge the difference that forms of production may make as a component of overall cost.

**Problem:** *You are a requesting party in a federal case, and you have made a timely, compliant and unambiguous written request for production of responsive information in native and near-
native forms. You have expressly requested that Microsoft Word documents be produced in their native .DOC or .DOCX formats. Your opponent instead produces Word documents to you as multiple .TIFF image files accompanied by a load file containing the extracted text from each document. When you object, your opponent counters that “this is what they always do” and that “TIFF plus load file is reasonably usable, so the Rules gave them the right to substitute TIFFs for natives.”

Assume that your opponent has produced 1,000 different Word documents which (for ease in making the calculation) are all exactly the same size as the native and converted file sizes for the file Always_and_Never.DOCX. Assume that none of the documents are privileged or required redaction. None are hash-matching duplicates of any other items produced.

You’ve contracted with an e-discovery service provider to load and host the documents produced so you can review and tag the documents for use in the case. The service provider charges by the gigabyte to ingest, process and host the data month-to-month. This is the applicable fee schedule:

To Ingest and Process Data Supplied:

<table>
<thead>
<tr>
<th>GB Range</th>
<th>Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 300 GB</td>
<td>$75.00 per GB</td>
</tr>
<tr>
<td>301 GB to 1 TB</td>
<td>$55.00 per GB</td>
</tr>
<tr>
<td>Greater than 1 TB</td>
<td>$40.00 per GB</td>
</tr>
</tbody>
</table>

Monthly Hosting Fee:

<table>
<thead>
<tr>
<th>GB Range</th>
<th>Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 300 GB</td>
<td>$23.00 per GB</td>
</tr>
<tr>
<td>301 GB to 1 TB</td>
<td>$20.00 per GB</td>
</tr>
<tr>
<td>Greater than 1 TB</td>
<td>$17.00 per GB</td>
</tr>
</tbody>
</table>

Any fraction of a gigabyte will be rounded up to a full gigabyte when calculating charges.

You intend to approach the Court to compel your opponent to produce the documents in the form you designated, and in addition to raising issues of utility, completeness and integrity, you want to determine whether the form produced to you will prove more expensive to ingest, process and host for the one-year period you expect to have the data online.

**Question:** If you accept the production in TIFF and load file, approximately how much more will it cost you over twelve months versus the same production in native forms?
How to Solve this Problem:

**Step 1: Normalize the file sizes.** Because the prices are quoted in gigabytes, you will want to express all data volumes in gigabytes, rather than as kilobytes or megabytes.

**Remember:** A kilobyte is one thousand bytes. A megabyte is one thousand kilobytes. A gigabyte is one thousand megabytes and a terabyte is one thousand gigabytes.

**Step 2: Calculate the cost of Native Production using normalized values:**

**Native Production:** One thousand files, each 18.3KB in size, is 18,300KB or 18.3MB. Because the service provider’s minimum charge is one gigabyte. The cost to ingest and host for one year would be:

\[
\text{Ingest and Process (1GB at \$75.00/GB)} + \text{Hosting (1GB at \$23.00/GB/month x 12 months)} = \$351.00
\]

**Step 3: Calculate the cost of TIFF and Text Load File Production using normalized values:**

**TIFF Plus Production:** One thousand files, each \((X) \text{ MB in size, is } (X) \text{ GB, where } (X) \text{ is the size of the file Always_and_Never.TIFF. We must also add the extracted text in the load file, which will be one thousand times } (Y) \text{ where } (Y) \text{ is the size of the the Always_and_Never.TXT file. Any fraction of a gigabyte should be rounded up to the next whole gigabyte. Consequently, the value } (Z) \text{ is the sum of } X \text{ plus } Y \text{ rounded up to the next whole gigabyte.}

\[
\text{Ingest and Process (ZGB at \$75.00/GB)} + \text{Hosting (Z GB at \$23.00/GB/month x 12 months)} = \$
\]

**Example calculation using hypothetical values:**

For example, if Always_and_Never.TIFF was 19MB in size and Always_and_Never.TXT was 57KB in size, the calculation would be:

\[
X = 1,000 \text{ (files) times 19MB} = 19GB
\]

\[
Y = 1,000 \text{ (text extractions) times 57KB} = 57MB = .057GB
\]

\[
Z = (19GB + .057GB) = 20GB \text{ (rounded up)}
\]

These values would make the calculation of the cost to ingest, process and host the TIFF Plus production:
Ingest and Process (20GB at $75.00/GB) + Hosting (20 GB at $23.00/GB/month x 12 months) = $7,020.00

The cost difference would be ($7,020.00 less $351.00) = $6,669.00.

Step 4: Calculate the difference using the actual file sizes obtained by your conversion of the file Always_and_Never.DOCX to TIFF and TXT.

What is the actual difference in cost comparing the native production to the TIFF plus TXT load file production?

Enter the actual difference here: $ ____________
Preparing for Meet and Confer

Federal Rule of Civil Procedure 26(f) requires parties to confer about preserving discoverable information and to develop a proposed discovery plan addressing discovery of electronically stored information and the form or forms in which it should be produced. This conference\textsuperscript{69}, and the overall exchange of information about electronic discovery, is called “meet and confer.”\textsuperscript{70}

Meet and confer is more a process than an event. Lay the foundation for a productive process by communicating your expectations. Send a letter to opposing counsel a week or two prior to each conference identifying the issues you expect to cover and sharing the questions you plan to ask.

E-discovery duties are reciprocal. At meet and confer, be prepared to answer many of the same questions you’ll pose. And while the focus will be on large data stores of ESI, don’t forget that even if your client has little electronic evidence, you must nonetheless act to preserve and produce it.

If you want client, technical or vendor representatives in attendance, say so. If you’re bringing a technical or vendor representative, tell them. Give a heads up on forms of production you’ll seek or are prepared to offer. Study up on any load file specification you want used and keywords to search, if only to let the other side know you’ve done your homework. True, your requests may be ignored or even ridiculed, but it’s not an empty exercise. A cardinal rule for electronic discovery, indeed for any discovery, is to tell your opponent what you seek or possess, plainly and clearly. They may show up empty-handed, but not because you failed to set the agenda.

\textsuperscript{69} The Fed. R. Civ. P. 26(f) conference must occur “as soon as practicable and in any event at least 21 days before a scheduling conference is held or a scheduling order is due under Rule 16(b).”

\textsuperscript{70} Hopson v. Mayor of Baltimore, 232 F.R.D. 228, 245 (D. Md. 2006) details some of counsel’s duties under Fed. R. Civ. P. 26(f): “[C]ounsel have a duty to take the initiative in meeting and conferring to plan for appropriate discovery of electronically stored information at the commencement of any case in which electronic records will be sought. … At a minimum, they should discuss: the type of information technology systems in use and the persons most knowledgeable in their operation; preservation of electronically stored information that may be relevant to the litigation; the scope of the electronic records sought (i.e., e-mail, voice mail, archived data, back-up or disaster recovery data, laptops, personal computers, PDA’s, deleted data) the format in which production will occur (will records be produced in “native” or searchable format, or image only; is metadata sought); whether the requesting party seeks to conduct any testing or sampling of the producing party’s IT system; the burdens and expenses that the producing party will face based on the Rule 26(b)(2) factors, and how they may be reduced (i.e. limiting the time period for which discovery is sought, limiting the amount of hours the producing party must spend searching, compiling and reviewing electronic records, using sampling to search, rather than searching all records, shifting to the producing party some of the production costs); the amount of pre-production privilege review that is reasonable for the producing party to undertake, and measures to preserve post-production assertion of privilege within a reasonable time; and any protective orders or confidentiality orders that should be in place regarding who may have access to information that is produced.”
The early, extensive attention to electronic evidence may nonplus lawyers accustomed to the pace of paper discovery. Electronic records are ubiquitous. They're more dynamic and perishable than their paper counterparts, require special tools and techniques to locate and process and implicate daunting volumes and multifarious formats. These differences necessitate immediate action and unfamiliar costs. Courts judge harshly those who shirk their electronic evidence obligations.

Questions for Meet and Confer

The following exemplar questions illustrate the types and varieties of matters discussed at meet and confer. They're neither exhaustive nor unique to any type of case, but are offered merely as talking points to stimulate discussion.

1. **What’s the case about?**

Relevance remains the polestar for discovery, no matter what form the evidence takes. The scope of preservation and production should reflect both claims and defenses. Pleadings only convey so much. Be sure the other side understands your theory of the case and the issues you believe should guide their retention and search.

2. **Who are the key players?**

Cases are still about people and what they did or didn’t say or do. Though there may be shared repositories and databases to discover, begin your quest for ESI by identifying the people whose conduct is at issue. These key players are custodians of ESI, so determine what devices and applications they use and target their relevant documents, application data and electronic communications. Too, determine whether assistants or secretaries served as proxies for key players in handling e-mail or other ESI.

Like so much in e-discovery, identification of key players should be a collaborative process, with the parties sharing the information needed for informed choices.

3. **What events and intervals are relevant?**

The sheer volume of ESI necessitates seeking sensible ways to isolate relevant information. Because the creation, modification, and access dates of electronic documents tend to be tracked, focusing on time periods and particular events helps identify relevant ESI, but only if you understand what the dates signify and when you can or can't rely on them. The Created Date of a document doesn't necessarily equate to when it was written. Neither does "accessed" always mean "used." For ESI, the “last modified” date tends to be the most reliable.
4. When do preservation duties begin and end?

The parties should seek common ground concerning when the preservation duty attached and whether there is a preservation duty going forward. The preservation obligation generally begins with an expectation of litigation, but the facts and issues dictate if there is a going forward obligation to preserve throughout the course of the litigation. Sometimes, events like plant explosions or corporate implosions define the endpoint for preservation, whereas a continuing tort or loss may require periodic preservation for months or years after the suit is filed. Even when a defendant’s preservation duty is fixed, a claimant’s ongoing damages may necessitate ongoing preservation.

5. What data are at greatest risk of alteration or destruction?

ESI is both tenacious and fragile. It’s hard to obliterate but easy to corrupt. Once lost or corrupted, ESI can be very costly or impossible to reconstruct. Focus first on fragile data, like storage media slated for reuse or messaging subject to automatic deletion, and insure its preservation. Address backup tape rotation intervals, disposal of legacy systems (e.g., obsolete systems headed for the junk heap), and re-tasking of machines associated with new and departing employees or replacement of aging hardware.

6. What steps have been or will be taken to preserve ESI?

Sadly, there are dinosaurs extant who believe all they have to reveal about ESI preservation is, “We’re doing what the law and the Rules require.” But that’s a risky tack, courting spoliation liability by denying you an opportunity to address problems before irreparable loss. More enlightened litigants see that reasonable disclosures serve to insulate them from sanctions for preservation errors.

7. What nonparties hold information that must be preserved?

ESI may reside with former employees, attorneys, agents, accountants, outside directors, Internet service providers, contractors, Cloud service providers, family members and other nonparties. Some of these non-parties may retain copies of information discarded by a party. Absent a reminder, litigants may focus on their own data stores and fail to take steps to preserve and produce data held by others over whom they have rights of direction or control.

8. What data require forensically sound preservation?

“Forensically sound” preservation of electronic media preserves, in a reliable and authenticable manner, an exact copy of all active and residual data, including remnants of deleted data residing in unallocated clusters and slack space. When there are issues of data loss,
destruction, alteration or theft, or when a computer is an instrumentality of loss or injury, computer forensics and attendant specialized preservation techniques may be required. Though skilled forensic examination can be expensive, forensically-sound preservation can cost less than $500 per system. So talk about the need for such efforts, and if your opponent won’t undertake them, consider whether you should force forensic preservation, even if you must bear the cost.

9. What metadata are relevant, and how will it be preserved, extracted and produced?

Metadata is evidence, typically stored electronically, that describes the characteristics, origins, usage and validity of other electronic evidence. There are all kinds of metadata found in various places in different forms. Some is supplied by the user, and some is created by the system. Some is crucial evidence, and some is just digital clutter. You will never face the question of whether a file has metadata—all active files do. Instead, the issues are what kinds of metadata exist, where it resides and whether it’s potentially relevant such that it must be preserved and produced. Understanding the difference—knowing what metadata exists and what evidentiary significance it holds—is an essential skill for attorneys dealing with electronic discovery.

The most important distinction is between application metadata and system metadata. The former is used by an application like Microsoft Word to embed tracked changes and commentary. Unless redacted, this data accompanies native production (that is, production in the form in which a file was created, used and stored by its associated application); but for imaged production, you’ll need to insure that application metadata is made visible before imaging or furnished in a useful form via a separate container called a “load file.”

System metadata is information like a file's name, size, location, and modification date that a computer’s file system uses to track and deploy stored data. Unlike application metadata, computers store system metadata outside the file. It’s information essential to searching and sorting voluminous data and therefore it should be routinely preserved and produced.

Try to get your opponent to agree on the metadata fields to be preserved and produced, and be sure your opponent understands the ways in which improper examination and collection methods corrupt metadata values. Also discuss how the parties will approach the redaction of metadata holding privileged content.

10. What are the parties’ data retention policies and practices?

A retention policy might fairly be called a destruction plan, and there’s always a gap—sometimes a chasm—between an ESI retention policy and reality. The more onerous the
policy, the greater ingenuity employees bring to its evasion to hang on to their e-mail and documents. Consequently, you can’t trust a statement that ESI doesn’t exist simply because a policy says it should be gone.

Telling examples are e-mail and backup tapes. When a corporate e-mail system imposes an onerous purge policy, employees find ways to store messages on, e.g., local hard drives, thumb drives and personal accounts. Gone from the e-mail server rarely means gone for good. Moreover, even companies that are diligent about rotating their backup tapes and that regularly overwrite old contents with new may retain complete sets of backup tapes at regular intervals. They also fail to discard obsolete tape formats when they adopt newer formats.

To meet their discovery obligations, the defendant may need to modify or suspend certain data retention practices. Discuss what they are doing and whether they will, as needed, agree to pull tapes from rotation or modify purge settings.

11. Are there legacy systems to be addressed?

Computers and servers tend to stick around even if they’ve fallen off the organization’s radar. That old laptop in someone’s drawer can serve as a time tunnel back to evidence thought long gone. You should discuss whether potentially relevant legacy systems exist and how they will be identified and processed. Likewise, you may need to address what happens when a key custodian departs. Will the system be re-assigned, and if so, what steps will be taken to preserve potentially relevant ESI?

12. What are the current and prior e-mail applications?

E-mail systems are Grand Central Station for ESI. Understanding the current e-mail system and other systems used in the relevant past is key to understanding where evidence resides and how it can be identified and preserved. On-premise corporate e-mail systems tend to split between the predominant Microsoft Exchange Server software tied to the Microsoft Outlook e-mail client on user’s machines and the less-encountered Lotus’ Domino mail server accessed by the Lotus Notes e-mail client application. Increasingly, companies dispense with maintaining physical systems altogether and deploy their e-mail systems online, “in the cloud.” Many companies now use Microsoft Office 365 and its virtualized version of the Exchange Server. A changeover from an old system to a new system, or even from an old e-mail client to a new one, can result in a large volume of “orphaned” e-mail on media that would not otherwise be ripe for search.
13. Are personal e-mail accounts and computer systems involved?

Those who work from home, out on the road or from abroad may use personal e-mail accounts for business or store relevant ESI on their home or laptop machines or other portable devices. Parties should address the potential for relevant ESI to reside on personal and portable machines and devices and agree upon steps to be taken to preserve and produce that data.

14. What electronic formats are common and in what anticipated volumes?

Making the right choices about how to preserve, search, produce and review ESI depends upon the forms and volume of data. Producing a Word document as a TIFF image may be acceptable where producing a native voice mail format as a TIFF is inconceivable. It’s difficult to designate suitable forms for production of ESI when you don’t know its native forms. Moreover, the tool you’ll employ to review millions of e-mails is likely much different than the tool you’ll use for thousands. If your opponent has no idea how much data they have or the forms it takes, encourage or compel them to use sampling of representative custodians to perform a “data biopsy” and gain insight into their collection.

15. How will we handle social networking, instant messaging and other challenging ESI?

Producing parties routinely ignore short-lived electronic evidence like social networking posts and instant messaging by acting too late to preserve it or deciding that the retention burden outweighs any benefit. *When it’s relevant*, will the other side archive texts, voice mail messages, social networking content, mobile device application content or a host of other potentially relevant ESI that’s often overlooked?

16. What relevant databases exist and how will their contents be discovered?

From R&D to HR and from finance to the factory floor, businesses run on databases. When they hold relevant evidence, you’ll need to know the platform (e.g., SQL, Oracle, SAP) and how the data’s structured (its “schema”) before proposing sensible ways to preserve and produce it. Options include generating standard reports, running agreed queries, exporting relevant data to standard delimited formats or even (in the very rare case) mirroring the entire contents to a functional environment.

Database discovery is challenging and contentious, so know what you need and articulate why and how you need it. Be prepared to propose reasonable solutions that won’t unduly disrupt operations.
17. Will paper documents be scanned, with what resolution, OCR and metadata?

Paper is still with us and ideally joins the deluge of ESI in ways that make it electronically searchable. Though parties are not obliged to convert paper to electronic forms, they commonly do so by scanning, coding and use of Optical Character Recognition (OCR). You’ll want to insure that paper documents are scanned so as to be legible and suited to OCR and are accompanied by information about their source (custodian, location, container, etc.) and logical unitization (i.e., folding and stapled and clipped groupings).

18. Are there privilege issues unique to ESI?

Discussing privilege at meet and confer entails more than just agreeing to return items that slip through the net via so-called “clawback agreements” or a Federal Rules of Evidence Rule 502 agreement or order. It’s important to surface practices that overreach. If the other side uses keywords to sidetrack potentially privileged ESI, are search terms absurdly overbroad? Simply because a document has the word “law” or “legal” in it or was copied to someone in the legal department doesn’t justify its languishing in privilege purgatory. When automated mechanisms replace professional judgment concerning the privileged character of ESI, those mechanisms must be closely scrutinized and challenged when flawed.

Asserting privilege is a privilege that should be narrowly construed to protect either genuinely confidential communications exchanged for the purpose of seeking or receiving legal counsel or the thinking and strategy of counsel. Moreover, even documents with privileged content may contain non-privileged material that should be parsed and produced. All the messages in a long thread aren’t necessarily privileged because a lawyer got copied on the last one.71

Electronic evidence presents unique privilege issues for litigants, in part because of the potential for application metadata (like documents comments and other collaboration features) to serve as communication tools. Comments and Track Changes aren’t fundamentally different from e-mails discussing suggested amendments to documents, yet the former tend not to be reviewed or produced by defendants. Instead, some parties will, e.g., convert Word documents to TIFF images, suppressing the embedded communications as if they never occurred so as to avoid having to review them for privilege. If these communications exist and may be relevant, you must work to insure this evidence is not ignored.

19. What search techniques will be used to identify responsive or privileged ESI?

Transparency of process is vitally important with respect to the mechanisms of automated search and filtering employed to identify or exclude information, yet opponents may resist sharing these details, characterizing it as work product. The terms and techniques facilitating an attorney’s assessment of a case are protected, but search and filtering mechanisms that effectively eliminate the exercise of attorney judgment by excluding data as irrelevant should be disclosed so that they may be tested and, if flawed, challenged. Likewise, if the producing party uses mechanized search to segregate data as privileged, the requesting party should be made privy to same in case it is inappropriately exclusive, though here, redaction may be appropriate to shield searches tending to reveal privileged information. Finally, use of advanced analytic techniques like predictive coding should be thoroughly explored to insure that the processes employed are well-understood and, as feasible, the sampling and thresholds are mutually acceptable.

20. If keyword searching is contemplated, can the parties agree on keywords?

If you’ve been to Las Vegas, you know Keno, that game where you pick the numbers, and if enough of your picks light up on the board, you win. Keyword searching ESI is like that. The other side has you pick keywords and then goes off somewhere to run them. Later, they tell you they looked through the matches and, sorry, you didn’t win. As a consolation prize, you may get the home game: a zillion jumbled images of non-searchable nonsense.

Perhaps because it performs so well in the regimented setting of online legal research, lawyers and judges invest too much confidence in keyword search. It’s a seductively simple proposition: pick the words most likely to uniquely appear in responsive documents and then review for relevance and privilege just those documents containing the key words. Thanks to, e.g., misspellings, acronyms, synonyms, IM-speak, noise words, OCR errors, indexing issues and the peculiar industry lexicons, keyword search performs far below most lawyers’ expectations, finding perhaps 20% of responsive material on first pass.72

Warts and all, keyword search remains the most common method employed to tackle large volumes of ESI, and a method still enjoying considerable favor with courts.

---

72 See, e.g., The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery (2007) (describing the famous Blair and Maron study, which demonstrated the significant gap between the assumptions of lawyers that they would find 75% of the total universe of relevant documents, versus the reality that they had in fact found only 20% of the total relevant documents in a 40,000 document collection).
Never allow your opponent to position keyword search as a single shot in the dark. You must be afforded the opportunity to use information gleaned from the first effort or subsequent efforts to narrow and target succeeding searches. The earliest searches are best used to acquaint both sides with the argot of the case. What shorthand references and acronyms did they use? Were products searched by their trade or technical names?

Collaborating on search terms is optimum, but a requesting party must be wary of an opponent who, despite enjoying superior access to and understanding of its own business data, abdicates its obligation to identify responsive information. Beware of an invitation to “give us your search terms” if the plan is to review only documents “hit” by your terms and ignore the rest. Also, insure that terms are tested on representative samples of ESI to insure that search tools and queries are performing as expected. Be especially wary of stop word exclusions and documents whose textual content was not extracted and indexed.

21. How will deduplication be handled, and will data be re-populated for production?

ESI, especially e-mail, is characterized by enormous repetition. A message may appear in the mail boxes of thousands of custodians or be replicated dozens or hundreds of times through periodic backup. Deduplication is the process by which identical items are reduced to a single instance for purposes of review. Deduplication can be vertical, meaning the elimination of duplicates within a single custodian’s collection, or horizontal, where identical items of multiple custodians are reduced to single instances.

Depending upon the review platform you employ, if production will be made on a custodial basis (person-by-person), it may be desirable to request re-population of content deduplicated horizontally so each custodian’s collection is complete. This will re-inject duplicates; however, each custodian’s collection will be complete, witness-by-witness.

22. What forms of production are offered or sought?

Notably, the 2006 Federal Rules amendments gave requesting parties the right to designate the form or forms in which ESI is to be produced. A responding party may object to producing the designated form or forms, but if the parties don’t subsequently agree and the court doesn’t order the use of particular forms, the responding party must produce ESI as it is ordinarily maintained or in a form that is reasonably usable. Moreover, responding parties may not simply dump undesignated forms on the requesting party, but must disclose the other forms
before making production so as to afford the requesting party the opportunity to ask the court to compel production in the designated form or forms. 73

Options for forms of production include native file format, near-native forms (e.g., individual e-mail messages in MSG or EML formats), imaged production (PDF or, more commonly, TIFF images accompanied by load files containing searchable text and metadata) and even paper printouts for very small collections. It is not necessary—and rarely advisable—to employ a single form of production for all items; instead, tailor the form to the data in a hybrid production. TIFF and load files may suffice for simple textual content like e-mail without attachments or word processed documents, but native forms are best for spreadsheets, documents with pertinent application metadata (comments and tracked changes) and social media content. Native forms are essential for rich media, like animated PowerPoint presentations or audio and video files. Quasi-native forms are well-suited to e-mail and database exports.

A requesting party uncertain of what he needs plays into the other side’s hands. You must be able to articulate both what you seek and the form in which you seek it. The native forms of ESI dictate the optimum forms for its production, but rarely is there just one option. The alternatives entail tradeoffs, typically sacrificing utility or searchability of electronic information to make it function more like paper documents. Before asking for anything, know how you’ll house, review and use it. That means “know your review platform.” 74 That is, know the needs

73 Fed. R. Civ. P. 34(b)  
74 If a question about your review platform gives you that deer-in-headlights look, you’re probably not ready for meet and confer. Even if you’re determined to look at every page of every item they produce, you’ll still need a system to view, search and manage electronic information. If you wait until the data start rolling in to pick your platform, you’re likely to get ESI in forms you can’t use, meaning you’ll have to expend time and money to convert them. Knowing your intended platform allows you to designate proper load file formats and determine if you can handle native production.

Choosing the right review platform for your practice requires understanding your work flow, your people, the way you’ll search ESI and the forms in which the ESI will be produced. You should not use native applications to review native production in e-discovery. Instead, a platform geared to review of ESI in native formats—one able to open the various types of data received without corrupting its content or metadata—should be employed. ESI can be like Russian nesting dolls in that a compressed backup file (.BKF) may hold an encrypted Outlook e-mail container (.PST) that houses a message transmitting a compressed archive (.ZIP) attachment containing an Adobe portable document (.PDF). Clearly, a review platform needs to be able to access the textual content of compressed and proprietary formats and drill down or “recurse” through all the nested levels.

There are many review platforms on the market, including the familiar Concordance and Summation applications, Internet-accessible hosted review environments like Relativity or iConect, and proprietary platforms marketed by e-discovery service providers touting more bells and whistles than a Mardi Gras parade.

Review platforms can be cost-prohibitive for some practitioners. If you don’t currently have one in-house, your case may warrant hiring a vendor offering a hosted platform suited to the ESI. When tight budgets make even that infeasible, employ whatever productivity tools you can cobble together on a shoestring. You may have to forego the richer content of native production in favor of paper-like forms such as Tagged Image File Format (TIFF) images because you can view them in a web browser.
and capabilities of the applications or tools you’ll employ to index, sort, search and access electronic evidence.

Finally, don’t let your opponent confuse the medium of production with the form of production. Telling you that the data is coming on a thumb drive tells you nothing about what data you’re getting.

23. How will you handle redaction of privileged or confidential content?

Defendants often seek to redact ESI in the way they once redacted paper documents: by blacking out text. To make that possible, ESI are converted to non-searchable TIFF images in a process that destroys electronic searchability. So after redaction, electronic searchability must be restored by using OCR to extract text from the TIFF image.

A TIFF-OCR redaction method works reasonably well for text documents, but it fails miserably applied to complex and dynamic documents like spreadsheets and databases. Unlike text, you can’t spell check numbers, so the inevitable errors introduced by OCR make it impossible to have confidence in numeric content or reliably search the data. Moreover, converting a spreadsheet to a TIFF image strips away its essential functionality by jettisoning the underlying formulae that distinguishes a spreadsheet from a table.

For common productivity applications like Adobe Acrobat and Microsoft Office, it’s increasingly feasible and cost-effective to redact natively so as to preserve the integrity and searchability of evidence; consequently, where it’s important to preserve the integrity and searchability of redacted documents, you should determine what redaction methods are contemplated and seek to agree upon methods best suited to the task. At all events, redaction tends to implicate a relatively small population of information items in a production; so, don’t let the preferred method of redaction adversely impact the form or forms of production employed for items not requiring redaction. That is, don’t let the redaction tail wag the production dog.

24. Will load files accompany document images, and how will they be populated?

Converting ESI to TIFF images strips the evidence of its electronic searchability and metadata. Accordingly, load files accompany TIFF image productions to hold searchable text and selected metadata. Load files are constructed of delimited text, meaning that values in each row of data follow a rigid sequence and are separated by characters like commas, tabs or quotation marks. Using load files entails negotiating their organization or specifying the content and the use of a structure geared to review software such as Summation, Concordance, Ringtail or Relativity.
25. How will the parties approach file naming and Bates numbering?

It’s common for file names to change to facilitate unique identification when ESI is processed for review and production. Assigned names may reflect, e.g., unique values derived from a data fingerprinting process called hashing or contain sequential control numbers tied to a project management database. Native productions don’t lend themselves to conventional paged formats, so aren’t suited to embossed Bates numbering on a page-by-page basis; however, this is no impediment to native production in that Bates numbers can serve as filenames for native files, with page numbers embossed on the items only when converted to paged formats for use in proceedings.

26. What ESI will be claimed as not reasonably accessible, and on what bases?

Pursuant to Rule 26(b)(2)(B) of the Federal Rules of Civil Procedure, a litigant must show good cause to discover ESI that is “not reasonably accessible,” but the burden of proving a claim of inaccessibility lies with the party resisting discovery. So, it’s important that your opponent identify the ESI it claims is not reasonably accessible and furnish sufficient information about that claim to enable you to gauge its merit.

The meet and confer is an opportune time to resolve inaccessibility claims without court intervention—to work out sampling protocols, cost sharing and filtering strategies—or when agreements can’t be reached, at least secure commitments that the disputed data will be preserved long enough to permit the court to resolve issues.

27. Can costs be minimized by shared providers, neutral experts or special masters?

Significant savings may flow from sharing costs of e-discovery service providers and online repositories, or by eliminating dueling experts in favor of a single neutral expert for thorny e-discovery issues or computer forensics. Additionally, referral of issues to a well-qualified ESI Special Master can afford the parties speedier resolution and more deliberate assessment of technical issues than a busy docket allows.

Endgame: Transparency of Process and Cooperation

Courts and commentators uniformly cite the necessity for transparency and cooperation in electronic discovery, but old habits die hard. Too many treat meet and confer as a perfunctory exercise, reluctant to offer a peek behind the curtain. Some are paying dearly for their intransigence, sanctioned for obstructive conduct or condemned to spend obscene sums chasing data that might never have been sought had there been communication and candor.
Others are paying attention and have begun to understand that candor and cooperation in e-discovery isn’t a sign of weakness, but a hallmark of professionalism.

The outsize cost and complexity of e-discovery will diminish as electronic records management improves and ESI procedures become standardized, but the meet and confer process is likely to endure and grow within federal and state procedure. Accordingly, learning to navigate meet and confer—to consistently ask the right questions and be ready with the right answers—is an essential advocacy skill.
Exercise 18: Meet and Confer

Students will form teams representing the plaintiff or defendant in a hypothetical case styled, *Lost Creek Engineering, LLC v. Keith Austin Weird and Artemis Energy Solutions, Inc.*, pending in a United States District Court. The nature of the case is described in Appendix A to this workbook, which also contains a chronology and case documents.

You will prepare for and engage in an FRCP Rule 26(f) meet and confer process with an opposing team. There will be a practice round against another team on Thursday evening and a final round against a different team on Friday morning. Teams should be prepared to answer questions that should be anticipated in a meet and confer and assess issues and information of importance to your client(s), most particularly on those points that must be addressed and reported to the Court pursuant to FRCP Rule 16 and Rule 26(f). Each side will be privy to information not known to their opponent that will influence how to proceed and the proper level of transparency and cooperation to offer and expect. In the final round on Friday, the teams will be observed by a sitting judge (assumed not to be present) and, when the judge directs, the teams will immediately appear at a scheduling conference (i.e., a hearing) before the judge where they will demonstrate their ability to present and explain the discovery plan and expertly and succinctly present unresolved issues to the court for resolution.

**Special Instructions**

Your team will receive a confidential plaintiff or defendant briefing. You are not to share the contents of this briefing with anyone other than your own team. You should not furnish or display same to a member of any other team nor should you look at an opponent’s briefing, if available to you. Acting in the best interests of your client(s) and consistent with your ethical duties, you may disclose information gleaned from the briefing in the meet and confer process only as legal requirements, good practice and sound strategy dictate.

The following provisions of the Federal Rules of Civil Procedure govern the conferences with your opponents and the Court. You should focus on the aspects of the process that bear on electronically stored information. You should not devote significant time to the merits of the action or to procedural matters that do not bear on e-discovery.
Rule 26. Duty to Disclose; General Provisions; Governing Discovery

... (f) Conference of the Parties; Planning for Discovery.

(1) Conference Timing. Except in a proceeding exempted from initial disclosure under Rule 26(a)(1)(B) or when the court orders otherwise, the parties must confer as soon as practicable—and in any event at least 21 days before a scheduling conference is to be held or a scheduling order is due under Rule 16(b).

(2) Conference Content; Parties' Responsibilities. In conferring, the parties must consider the nature and basis of their claims and defenses and the possibilities for promptly settling or resolving the case; make or arrange for the disclosures required by Rule 26(a)(1); discuss any issues about preserving discoverable information; and develop a proposed discovery plan. The attorneys of record and all unrepresented parties that have appeared in the case are jointly responsible for arranging the conference, for attempting in good faith to agree on the proposed discovery plan, and for submitting to the court within 14 days after the conference a written report outlining the plan. The court may order the parties or attorneys to attend the conference in person.

(3) Discovery Plan. A discovery plan must state the parties' views and proposals on:

(A) what changes should be made in the timing, form, or requirement for disclosures under Rule 26(a), including a statement of when initial disclosures were made or will be made;

(B) the subjects on which discovery may be needed, when discovery should be completed, and whether discovery should be conducted in phases or be limited to or focused on particular issues;

(C) any issues about disclosure, discovery, or preservation of electronically stored information, including the form or forms in which it should be produced;

(D) any issues about claims of privilege or of protection as trial-preparation materials, including—if the parties agree on a procedure to assert these claims after production—whether to ask the court to include their agreement in an order under Federal Rule of Evidence 502;

(E) what changes should be made in the limitations on discovery imposed under these rules or by local rule, and what other limitations should be imposed; and

(F) any other orders that the court should issue under Rule 26(c) or under Rule 16(b) and (c).
Rule 16. Pretrial Conferences; Scheduling; Management

(a) Purposes of a Pretrial Conference. In any action, the court may order the attorneys and any unrepresented parties to appear for one or more pretrial conferences for such purposes as:

(1) expediting disposition of the action;
(2) establishing early and continuing control so that the case will not be protracted because of lack of management;
(3) discouraging wasteful pretrial activities;
(4) improving the quality of the trial through more thorough preparation; and
(5) facilitating settlement.

(b) Scheduling.

(1) Scheduling Order. Except in categories of actions exempted by local rule, the district judge—or a magistrate judge when authorized by local rule—must issue a scheduling order:
(A) after receiving the parties’ report under Rule 26(f); or
(B) after consulting with the parties’ attorneys and any unrepresented parties at a scheduling conference.

(2) Time to Issue. The judge must issue the scheduling order as soon as practicable, but unless the judge finds good cause for delay, the judge must issue it within the earlier of 90 days after any defendant has been served with the complaint or 60 days after any defendant has appeared.

(3) Contents of the Order.

(A) Required Contents. The scheduling order must limit the time to join other parties, amend the pleadings, complete discovery, and file motions.

(B) Permitted Contents. The scheduling order may:
(i) modify the timing of disclosures under Rules 26(a) and 26(e)(1);
(ii) modify the extent of discovery;
(iii) provide for disclosure, discovery, or preservation of electronically stored information;
(iv) include any agreements the parties reach for asserting claims of privilege or of protection as trial-preparation material after information is produced, including agreements reached under Federal Rule of Evidence 502;

(v) direct that before moving for an order relating to discovery, the movant must request a conference with the court;

(vi) set dates for pretrial conferences and for trial; and

(vii) include other appropriate matters.

(4) **Modifying a Schedule.** A schedule may be modified only for good cause and with the judge's consent.

....
All of the events and persons described in this hypothetical scenario are fictional. Any resemblance to persons, living or dead, or to business entities is purely coincidental.

This hypothetical case concerns the alleged misappropriation of intellectual property by a senior design engineer at an engineering company. The engineer, Keith Austin Weird, worked for Lost Creek Engineering, LLC for 15 years, rising to the position of Assistant Vice-President of Engineering. Weird led the design and development of Lost Creek’s very profitable Arnold™ line of intelligent pipeline pigs, as well as a yet-to-be-introduced line of next generation products codenamed “When Pigs Fly.”

Pigs, in the context of pipelines, are devices inserted into pipelines that travel with the flowing content for the purpose of conducting inspection, maintenance, product separation and other functions. Pipeline pigs must operate under conditions of high pressure, extreme temperatures and highly corrosive conditions. Intelligent or “smart” pigs are sophisticated robots that, until now, have been required to operate autonomously because the radio-blocking “Faraday cage” character of steel pipelines and the enormous distances traversed made it infeasible for pigs to communicate with remote operators or GPS satellites.

Lost Creek’s “When Pigs Fly” innovation was the pairing of its smart pigs with an accompanying drone aircraft outside the pipeline. The innovation employs proprietary technology to enable high-bandwidth, multichannel ultrasonic communications between pig and drone, allowing a distant operator to see real time data and video from the pig, obtain precise GPS coordinates and remotely control the pig. Precise location data means that repair crews operate more efficiently and at lower cost. Real time remote control permits complex repairs to be accomplished without the risk and cost of dispatching crews and heavy equipment to distant work sites.

Weird was hired by former Lost Creek V.P. of Engineering and Development, Montgomery Bonnell in 2000. Weird reported directly to Bonnell for the decade that both worked together at Lost Creek. The two are close friends, and their families frequently socialize outside of work. In 2010, Bonnell left Lost Creek to found Artemis Energy Solutions, Inc. in Houston. Artemis manufactures and sells pipeline telemetry products to the energy sector. Weird sought to be
considered for Bonnell’s position, but was told he was too valuable in his current position and encouraged to acquire some managerial seasoning. When an outsider was brought in to replace Bonnell, Weird was assured by the CEO that his desire to advance would not be forgotten. Bonnell’s replacement left Lost Creek at the start of 2015, and Weird learned that management contacted a headhunter to fill the position.

With no promotion forthcoming, Weird resigned from Lost Creek on August 14, 2015. He gave two weeks’ notice and noted that, now that his kids were in college, he was heading to Houston to work for his old friend, Monty Bonnell, at Artemis Energy Solutions, Inc. Weird participated in a required exit interview, confirmed his familiarity with all Lost Creek polices impacting departing employees, and received a generous severance package to resolve unused vacation time and other benefits. Weird’s last day at Lost Creek was August 28, 2015, and he took two weeks off before starting at Artemis. Weird joined Artemis as its Executive VP of Technology.

On December 21, 2015, Lost Creek’s outside counsel, Lamar Street, sent letters to Weird and Bonnell invoking the Non-Disclosure Agreement and Covenant Not to Compete Weird signed when first hired by Lost Creek. Lost Creek demanded that Weird cease work for Artemis on anything involving pipeline pigs or telemetry. The letter to Weird also sought return of Weird’s Lost Creek laptop and access to all of Weird’s personal computers, digital media and e-mail accounts for the purpose of conducting an examination to assess compliance.

Shortly after Lost Creek distributed year-end bonuses on December 23, 2015, three Lost Creek engineers, Percy Pennybacker, Claudia Johnson and Barton Springs, tendered their resignations. All had worked under Weird at Lost Creek in the development and testing of intelligent pipeline pigs. All joined Artemis and once more report to Weird.

In February of 2016, Artemis’ internal SharePoint newsletter announced that the company would be introducing the AirHog™ line of sophisticated intelligent pipeline drone pigs that, by the description of their capabilities, would mirror the capabilities of Lost Creek’s yet-to-be-introduced When Pigs Fly technology. The article offered rosy financial projections for the new product line, prompting a blizzard of Tweets and texts between Artemis employees, Lost Creek employees and industry insiders.

On March 1, 2016, Lost Creek filed suit against Weird and Artemis in the Western District of Texas seeking injunctive relief and damages on seven counts:

Count 1 – Breaches of Trade Secret Agreement and Covenant Not to Compete
Count 2 – Unfair Competition by Misappropriation
Count 3 – Tortious Conversion
Count 4 – Common Law Misappropriation of Trade Secrets
Count 5 – Tortious Interference with M-I’s Employment Contracts
Count 6 – Breach of Fiduciary Duty
Count 7 – Civil Conspiracy

The Defendants answered, asserting various affirmative defenses.

Lost Creek has been in business for 40 years. It is headquartered in Austin, Texas and maintains manufacturing sales and service centers in China, Australia and Europe, as well as representatives and technicians in more than 20 countries. Lost Creek is a closely-held company that employs over 400 people, including 40+ persons in its Product Development and Engineering Division. Its sales and earnings figures are not made public.

Artemis Energy Solutions, Inc. was formed in 2010 and is headquartered in Houston, Texas. Artemis employed 150 people as of April 15, 2016, and projected gross annual sales of approximately $75 million for 2016 based on first quarter results. In May of 2016, Artemis was acquired by Prytania Oil, S.A., a conglomerate headquartered in Greece, and Artemis became a wholly-owned foreign subsidiary of Prytania Oil, S.A.
Timeline of Events

September 1, 2000: Keith Austin Weird hired by Lost Creek; executes Non-Disclosure Agreement and Covenant Not to Compete

August 12, 2015: Weird receives offer letter from Artemis and copies Lost Creek data to an external hard disk drive

August 14, 2015: Weird tenders his resignation to Lost Creek

August 28, 2015: Weird’s last day at Lost Creek; exit interview

August 31 - September 11, 2015: Weird on vacation

September 14, 2015: Weird’s first day at Artemis

December 2015: Three Lost Creek engineering employees quit to join Artemis

December 21, 2015: Demand for return of Weird’s Lost Creek laptop and to inspect his e-mail, home systems, hard drives and thumb drives

February 1, 2016: Artemis announces forthcoming AirHog™ product line

March 1, 2016: Original Complaint filed

March 10, 2016: Original Answer filed

March 15, 2016: Amended Complaint Filed

March 18, 2016: Amended Answer filed

April 1, 2016: Agreed Temporary Injunction entered

May 15, 2016: Prytania Oil, S.A. acquires all shares in Artemis
The undersigned represents the legal interests of Lost Creek Engineering, LLC ("Lost Creek" or the "Company"). As you know, in connection with your employment with Lost Creek, you were given specialized training and were provided with certain of the Company's confidential, proprietary, and trade secret information. You expressly acknowledged this in Non-Disclosure Agreement and Covenant Not to Compete (the "Agreement"). A copy of the Agreement is enclosed for your reference.

Additionally, your contract of employment includes an agreement to refrain from working for a competitive business following the termination of your employment from Lost Creek. In the Agreement you promised that, for a period of two (2) years following your termination from Lost Creek, you would not engage in or work for any business in direct competition with Lost Creek by manufacturing and/or selling intelligent pipeline pigs that resemble or imitate the pipeline pigs manufactured and sold by Lost Creek. See Agreement at 1.

In your letter of resignation dated August 14, 2015, you indicated that you would be taking a position with Artemis Energy Solutions, Inc. as Technology Director-Pipeline Products. Although, Lost Creek does not consider Artemis to be directly competitive with its interests. Any work by you in support of the design and production of intelligent pipeline pigs is in direct competition with Lost Creek and in direct violation of the Agreement. As we now understand that your work with Artemis will be in research and development in remote sensing pipeline repair devices, a technical knowledge that you gained exclusively during your tenure at Lost Creek, the purpose of this correspondence is to notify you of your breach of the Agreement and demand that you cease your intent to continue employment with Artemis and refrain from doing so for a period of two (2) years. We also remind you that your agreement to protect confidential information that
belongs to Lost Creek is not limited in any timeframe and is your obligation regardless of employment status.

On behalf of Lost Creek, we demand immediate return of all files, materials, information, technology or other property owned by Lost Creek which may be in your possession. To be assured that you have complied with this request, we request that you deliver your Lost Creek laptop, any home computer(s) and any external drives and thumb drives to our forensic examiner (see attached business card) for forensic review of the hard drives and external drives to assure that no confidential information or property of Lost Creek resides on any drive. We further request that you make the contents of any e-mail or webmail accounts you have used within the last two (2) years available to us for inspection and copying. We will also seek confirmation that you have not distributed or transferred any such information to any third party including Artemis Energy Solutions, Inc. or any other manufacturer in the pipeline pig industry. Lost Creek will withhold the six (6) month severance pay provided in your Agreement pending compliance with this request.

In further effort to assure compliance with these post-employment requirements of you, Lost Creek has asked that you complete and sign the enclosed verification which confirms your representations that you do not have any information which could be considered confidential information belonging to Lost Creek.

Know that Lost Creek must and will protect its legal interests. Failure to immediately cease your employment with Artemis Energy Solutions, Inc. and provide the undersigned with satisfactory notice thereof will require the Company to take action to protect its legal interests. Such action will include the immediate imposition of suit against you to enforce the Agreement. In addition to the actual damages caused by your breach of the Agreement, Lost Creek will seek recovery of its attorneys' fees, costs, and interest. Please provide me with the requisite notice of termination of employment with Artemis Energy Solutions, Inc. at your earliest convenience and evidence of your compliance with the request that you deliver your computers to our forensic examiner.

Very Truly Yours,

Lamar Street
NON-DISCLOSURE AGREEMENT AND COVENANT NOT TO COMPETE

Lost Creek Engineering, LLC ("Lost Creek") hereby promises that, upon Keith Austin Weird's acceptance of employment with Lost Creek, Lost Creek will provide Keith Austin Weird with specialized training unique to it and not otherwise available in the industry or elsewhere. Further, Lost Creek promises to provide Keith Austin Weird with information it holds as confidential, as well as certain trade secret information relating to intelligent pipeline pigs designed, manufactured and sold by Lost Creek, including access to certain privileged materials.

During the term of employment and without limitation thereafter, Keith Austin Weird hereby covenants and agrees to keep strictly confidential all knowledge to which he gains by virtue of his employment with Lost Creek. This includes all trade secrets, business practices, finances, documents, blueprints, market data, other intellectual property and other confidential information. Keith Austin Weird agrees not to disclose the above mentioned confidential information, directly or indirectly to any other person, company or corporation, or use it for his own benefit. Keith Austin Weird agrees that he will only use the confidential information as an employee of Lost Creek.

All confidential or trade secret information relating to the business of Lost Creek which Keith Austin Weird shall develop, conceive, produce, construct or observe during his employment with Lost Creek shall remain the sole property of Lost Creek.

Keith Austin Weird further agrees that upon termination of his employment, Keith Austin Weird will surrender and deliver to Lost Creek all confidential information, including but not limited to work papers, books, records, and data of every kind relating to or in connection with Lost Creek.

Keith Austin Weird agrees, upon termination of employment with Lost Creek and for a period of two (2) years thereafter, Keith Austin weird will not directly or indirectly engage in any business or work for any business which is in direct competition with Lost Creek by manufacturing and/or selling pipeline pigs that resemble or imitate the pipeline pigs manufactured and sold by Lost Creek. Keith Austin Weird agrees that this paragraph prohibits him from accepting employment on a worldwide basis with any pipeline pig manufacturer for the two (2) year period.
Executed this 1st day of September, 2000

[Signature]

9/1/2000

Today's Date

The State of Texas  

County of Travis  

BEFORE ME the undersigned authority, on this day personally appeared Keith A. Weird known to me to be the person whose name is subscribed to the foregoing instrument, and acknowledged to me that he had executed same in the capacities and for the purposes and consideration therein expressed.

GIVEN UNDER MY HAND AND SEAL OF OFFICE THIS 1st day of September 2000

JANE DOE  
Notary Public, State of Texas  
Commission Expires 05-18-2001  

Notary Public in the State of Texas
VERIFICATION

My name is Keith Austin Weird. I have been employed with Lost Creek Engineering, LLC ("Lost Creek") as Vice-president of Engineering and later Chief of Engineering since September 1, 2000. I have resigned from employment with Lost Creek effective August 28, 2015. In connection with my resignation, I have been asked to represent and warrant that I am in compliance with certain agreements related to my employment. Accordingly, I represent and warrant that:

I am aware of my obligations under that certain agreement dated September 1, 2000 entitled Non-Disclosure Agreement and Covenant Not to Compete (the "Agreement") and agree to comply with my obligations under the Agreement to the fullest extent possible. I understand and agree that confidential information and trade secrets includes all trade secrets, customer and vendor information, business practices, finances documents, blueprints, market data, other intellectual property relating to Lost Creek’s work in the Pipeline pig industry, including remote sensing pipeline repair devices. I acknowledge that all information regarding remote sensing pipeline repair devices I have has been gained during my tenure with Lost Creek. I have not removed any confidential information or trade secrets from Lost Creek at any time during my employment. If I have any confidential information or trade secrets in my possession in written or electronic form, I will return it to Lost Creek immediately and no later than Friday, December 28, 2015.

I have not transferred any confidential information or trade secrets to any third party prior to my departure from Lost Creek. I agree to provide any computer and all external drives or devices, including jump drives, in my possession or use at home or elsewhere to Lost Creek’s designated agent for forensic review on or before December 28, 2015 or at such time as Lost Creek directs for the purpose of verifying removal of all confidential information belonging to Lost Creek from such computer. I further consent to allow Lost Creek’s designated agent to access and copy any personal e-mail or webmail account I have used for the last two (2) years.

Date: ____________________  ________________________________

______________________________________________
Keith Austin Weird
December 21, 2015

Montgomery Bonnell
Chief Executive Officer
Artemis Energy Solutions, Inc.
One Big Oil Boulevard
Houston, TX 77041

RE: Lost Creek Engineering, LLC.

Dear Mr. Bonnell:

We are counsel to Lost Creek Engineering, LLC ("Lost Creek"). We have been apprised of the fact that Keith Austin Weird has been offered employment with Artemis Energy Solutions, Inc. or one of its affiliates ("Artemis"). Lost Creek has recently learned that Mr. Weird's employment may involve research and development of intelligent pipeline pigs and/or remote sensing repair tools. If so, Mr. Weird would be performing the same (if not identical) services for Artemis as he performed for Lost Creek. We are writing, in part, to give you notice that Mr. Weird is subject to a prohibition from employment with a competitor of Lost Creek. A copy of Mr. Weird's agreement with Lost Creek is enclosed for your review. We are concerned that Mr. Weird's employment with Artemis may be in violation of the non-competition agreement and request your assistance in assuring his compliance with it.

Lost Creek is further concerned with Mr. Weird's compliance with his agreement to protect confidential information belonging to Lost Creek. It is possible that he possesses confidential information from Lost Creek' files and may inadvertently use such information in connection with his employment with Artemis. Mr. Weird's entire knowledge regarding the intelligent pipeline pig industry has been gained during his employment with Lost Creek and we believe that, even with the best of intentions, it would be impossible for him to work in research and development regarding pipeline pigs without using Lost Creek's confidential information in violation of his
agreement to protect it. We therefore suggest to you that any employment of Mr. Weird which is in violation of his non-compete agreement or work performed by Mr. Weird and which may cause him to disclose confidential information belonging to Lost Creeks could result in legal action on behalf of Lost Creek. We trust that Artemis will work with Lost Creek to assure that there are no violations of the agreements or of other laws.

Please contact me if you have any questions.

Very Truly Yours,

Lamar Street
IN THE UNITED STATES DISTRICT COURT  
WESTERN DISTRICT OF TEXAS  

LOST CREEK ENGINEERING, L.L.C.  

§  

§  

§  

V.  

§  

§  

§  

KEITH AUSTIN WEIRD,  
and  

ARETEMIS ENERGY SOLUTIONS, INC.  

§  

§  

§  

§  

§  

§  

§  

AMENDED COMPLAINT  

TO THE HONORABLE JUDGE OF SAID COURT:  
COMES NOW, Lost Creek Engineering, L.L.C., hereinafter referred to as Plaintiff or “Lost Creek,” complaining of Keith Austin Weird and Artemis Energy Solutions, Inc. (“Artemis”), hereinafter referred to as Defendants, and for cause of action would respectfully show unto the Court and jury as follows:  

I. PARTIES  
1. Plaintiff is a corporation with an office in Travis County, Texas, and which has authority to do business in the State of Texas.  

2. Defendant Keith Austin Weird has been served and answered.  

3. Defendant, Artemis Energy Solutions, Inc., (“Artemis”) is a foreign corporation doing business in Texas and has been served and answered.  

II. VENUE AND JURISDICTION  
4. This Court has federal question and supplemental jurisdiction pursuant to 28 U.S.C. § 1331, 1441, 1367 and 18 U.S.C. § 1030.  

5. Venue is proper in the Western District of Texas because Weird resides in Travis County, Texas, and because a substantial part of the events or omissions giving rise to the claims plead below occurred in Travis County, Texas.  

III. FACTUAL BACKGROUND  
6. Lost Creek Engineering is a manufacturer of specialized tools for the inspection, maintenance and repair of petroleum and natural gas pipelines. Sophisticated and sensitive in-line inspection (ILI) tools travel through the pipe and measure and record irregularities that may
represent corrosion, cracks, laminations, deformations or other defects. Lost Creek is a world leader in the design, development and sale of pipeline smart pigs, robots designed to pass through pipelines performing specialized tasks in highly challenging environments. Lost Creek’s Arnold™ line of smart pigs employ proprietary state-of-the-art magnetic flux and ultrasonic sensing devices and high-definition imagery in ways that uniquely distinguish Lost Creek’s products in the marketplace.

7. Typically, smart pigs are inserted into the pipeline at a location, such as a valve or pump station, that has a special configuration of pipes and valves where the tool can be loaded into a receiver, the receiver can be closed and sealed, and the flow of the pipeline product can be directed to launch the tool into the main line of the pipeline. A similar setup is located downstream, where the tool is directed out of the main line into a receiver, the tool is removed, and the recorded data retrieved for analysis and reporting. Historically, smart pigs have been required to operate autonomously because the radio-blocking “Faraday cage” character of steel pipelines and the enormous distances traversed made it infeasible for pigs to communicate with remote operators or GPS satellites.

8. In utmost secrecy and through its investment of large sums of time and money, Lost Creek developed a unique and innovative technology to enable remote control and geolocation of the next generation of smart pig technology. Lost Creek’s Project When Pigs Fly” (WPF) innovation was the pairing of its smart pigs with an accompanying drone aircraft outside the pipeline. The innovation employs proprietary technology to enable high-bandwidth, multichannel ultrasonic communications between pig and drone, allowing a distant operator to see real time data and video from the pig, obtain precise GPS coordinates and remotely control the pig. Precise location data means that repair crews operate more efficiently and at lower cost. Real time remote control permits complex repairs to be accomplished without the risk and cost of dispatching crews and heavy equipment to distant work sites.

9. The design of Lost Creek’s WPF of smart pigs has been a time-consuming and expensive process. Lost Creek continually tests, researches and improves the components, materials, designs and manufacturing processes of its products. It has taken years of field tests, experiments, research and development for Lost Creek to develop the unique technologies it is poised to market to customers. There are specific design characteristics of Lost Creek’s smart pigs that are not used by other smart pig manufacturers and are not found in the open market. Such unique design characteristics include the following: (1) high-bandwidth, multichannel ultrasonic communications hardware, circuits and software; (2) Drone control and synchronization programming; (3) image and data compression algorithms; and (4) associated tools for inspection, optimization, deployment and operation of WPF drone/pig pairs.
10. These unique design characteristics were discovered and innovated by Lost Creek’s engineers over the past ten years through testing, research and experience. It is these design characteristics that differentiate Lost Creek’s smart pigs from other smart pigs on the market.

11. In order to design, test and (ultimately) manufacture smart pigs with WPF capabilities for its customers, Lost Creek uses specialized designs, test mechanisms, source code and algorithms (“WPF Proprietary Technology”). The information comprising Lost Creek’s WPF Proprietary Technology derive from and is the product of many years of experience, the labor of dozens of Lost Creek’s skilled employees, and millions of dollars invested by Lost Creek in research, testing, innovation and application. A competitor in possession of Lost Creek’s WPF Proprietary Technology would have the ability to develop products and compete with Lost Creek without expending the time, energy, and resources that Lost Creek expended to develop its unique products and technology. The information comprising Lost Creek’s WPF Proprietary Technology (e.g., specific formulas, designs, dimensions, safety factors, tolerances, programming source code, etc.) is not legitimately known outside of Lost Creek and provides a competitive advantage to Lost Creek in the marketplace.

12. Lost Creek has taken great care to ensure that the custom design features of its products and manufacturing processes are kept confidential and remain a trade secret. Lost Creek’s designs, testing, algorithms and other details of Lost Creek’s custom features cannot be found in the open market and are not available to competitors to view or reverse engineer. Lost Creek’s WPF Proprietary Technology is only accessible to a limited number of Lost Creek employees and are protected from disclosure through the compulsory use of access cards, usernames and passwords required to access the information. Furthermore, each Lost Creek employee that works with the WPF Proprietary Technology is required to sign a confidentiality agreement protecting such information from disclosure. As such, Lost Creek’s WPF Proprietary Technology is a trade secret of Lost Creek’s business.

13. Weird executed and agreed to his Non-Disclosure Agreement and Covenant Not to Compete (NDA/CNC) on September 1, 2000. Pursuant to the NDA/CNC, Weird agreed that upon termination of his employment with Plaintiff that he would maintain the confidentiality of Plaintiff’s technology, trade secrets and proprietary and confidential information. Weird also agreed not to compete against Plaintiff for two years after such termination of employment with Plaintiff and to refrain from certain activities in competition against Plaintiff, such as providing the same or similar function with a competitor as they provided to Lost Creek.

14. Defendant Keith Austin Weird was a long-time, trusted employee of Lost Creek. Weird worked for Lost Creek as an engineer for over fifteen years in its offices in Austin, Travis County, Texas. He was ultimately promoted to the position of Assistant Vice President of Engineering. Weird had duties and obligations to protect Lost Creek’s trade secrets and other confidential proprietary information from disclosure.
15. When he began employment with Lost Creek, Weird signed an NDA/CNC providing:

"During the term of employment and without limitation thereafter, Keith Austin Weird hereby covenants and agrees to keep strictly confidential all knowledge to which he gains by virtue of his employment with Lost Creek. This includes all trade secrets, business practices, finances, documents, blueprints, market data, other intellectual property and other confidential information. Keith Austin Weird agrees not to disclose the above mentioned confidential information, directly or indirectly to any other person, company or corporation, or use it for his own benefit. Keith Austin Weird agrees that he will only use the confidential information as an employee of Lost Creek.

All confidential or trade secret information relating to the business of Lost Creek which Keith Austin Weird shall develop, conceive, produce, construct or observe during his employment with Lost Creek shall remain the sole property of Lost Creek.

Keith Austin Weird further agrees that upon termination of his employment, Keith Austin Weird will surrender and deliver to Lost Creek all confidential information, including but not limited to work papers, books, records, and data of every kind relating to or in connection with Lost Creek.

Keith Austin Weird agrees, upon termination of employment with Lost Creek and for a period of two (2) years thereafter, Keith Austin weird will not directly or indirectly engage in any business or work for any business which is in direct competition with Lost Creek by manufacturing and/or selling pipeline pigs that resemble or imitate the pipeline pigs manufactured and sold by Lost Creek. Keith Austin Weird agrees that this paragraph prohibits him from accepting employment on a worldwide basis with any pipeline pig manufacturer for the two (2) year period."

16. During Weird's employment with Lost Creek, he worked with other Lost Creek engineers to develop the unique WPF Proprietary Technology. As a Lost Creek employee, Weird was involved in the research, development, calculations, drawings, testing and design of Lost Creek's products. Through his work for Lost Creek, Weird had knowledge of and access to research and designs, to the technical aspects of Lost Creek’s products and to the applications in which Lost Creek’s products function.

17. On August 17, 2015, Weird received a written offer of employment by e-mail from Montgomery Bonnell, CEO of Artemis and a former Vice-President of Lost Creek who hired and supervised Weird beginning in 2000 until Bonnell’s departure in 2010.

18. On August 17, 2015, Weird connected an external Western Digital My Passport hard drive to his Lost Creek laptop computer and downloaded almost thirty gigabytes of data comprising thousands of Lost Creek’s confidential business documents and trade secrets. Included
among this material were the complete contents of Weird’s “Documents” folder holding WPF Proprietary Technology. Also on August 17, 2015, Weird connected one or more USB thumb drives to his Lost Creek laptop.

19. On August 20, 2015, Weird submitted his resignation letter to Lost Creek, effective August 31. In his resignation letter, Weird advised Lost Creek that he would be assuming a position with Artemis Energy Solutions, Inc. (“Artemis”) as Technology Director-Pipeline Products. At the time of his resignation, Weird advised Lost Creek that prior to his departure, he would "return any and all confidential material belonging to Lost Creek that is in [his] possession.”

20. Upon information and belief, CEO Montgomery Bonnell and other Artemis officers or employees induced Weird to misappropriate Lost Creek’s confidential information and trade secrets for use in Artemis’ business operations.

21. Following Weird’s departure, Lost Creek discovered that Weird transferred numerous emails containing confidential and trade secret information to his personal webmail account.

22. On August 31, 2015, Weird participated in an exit interview wherein he was instructed to return any confidential business or trade secret information. Weird claimed he did not have any such information. When asked to return his Lost Creek laptop computer, Weird stated that he had left it at his home and promised to return it at a later date. Despite repeated requests that he do so, Weird has not returned his Lost Creek laptop. Weird has further declined to permit inspection of his webmail and has failed to respond to a written demand that he make his personal and Artemis computers, phones, tablets and data storage devices available for inspection.

23. Since Weird’s departure, Artemis has hired three former Lost Creek engineering employees, Percy Pennybacker, Claudia Johnson and Barton Springs, who worked on development and testing of Lost Creek’s WPF smart pig.

24. It is clear that Artemis targeted Lost Creek to poach its employees to start a smart pig division and begin manufacturing smart pigs in direct competition with Lost Creek. Artemis CEO, Montgomery Bonnell, approached Weird and, on information and belief, other Lost Creek employees with offers of employment and inducements of bonuses. Since Artemis had no smart pig division nor a smart pig product, hiring Lost Creek engineers was the shortest route to market.

25. On information and belief, Artemis began aggressively pursuing development of a WPF-like smart pig product line approximately six months before Weird was hired, but encountered difficulties due to the complexity of the complex technological challenges resolved by use of Lost Creek’s WPF technology. Weird was hired by Artemis to gain access to Lost Creek’s WPF Proprietary Technology as it enabled Artemis to develop competing products without created
expending the time and resources required to develop competing products through research and testing.

26. In February 2016, Artemis distributed a newsletter announcing that it would be expanding its product offerings to feature a new line of AirHog™ drone-paired, remote-controlled pipeline smart pigs. Weird was identified as leading the effort to bring the new products to market. Prior to Weird’s employment with Artemis, Artemis did not manufacture or sell any type of smart pig products that competed with Lost Creek’s products, let alone any product with the innovative and sophisticated features of Lost Creek’s WPF Proprietary Technology.

27. On information and belief, Artemis has contracted with existing clients of Lost Creek for the sale of AirHog™ products that imitate or resemble the WPF smart pigs developed by Lost Creek. Weird and Artemis have further applied for a patent on features of the design of the AirHog™ Remote-Controlled Pipeline Smart Pig. It is implausible that Artemis, lacking experience in the design and manufacture of smart pig products could design, develop, manufacture, patent and sell such products in less than eighteen months without unauthorized use of the WPF Proprietary Technology developed by Lost Creek.

IV. APPLICATION FOR INJUNCTIVE RELIEF

28. All previous paragraphs are incorporated herein.

29. Lost Creek requests a Permanent Injunction that Defendants, and each of their agents, servants, representatives, and all other persons or entities in active concert or participation with Defendants who receive actual notice of this Order by personal service or otherwise be and hereby are enjoined as follows:

a. Defendants are restrained from violating the Non-Disclosure Agreement and Covenant Not to Compete entered into between Lost Creek and Weird or participating in the violation of said NDA/CNC;

b. Defendants are ordered to return to Lost Creek, and to cease and desist from using, any Lost Creek proprietary documents, electronic files or other property, including but not limited to Lost Creek’s WPF Proprietary Technology or any Artemis document that uses Lost Creek’s information;

c. Defendants are restrained from altering or deleting any electronic files on their personal or work computers, mobile devices, PDs, smart phones, webmail accounts, online storage repositories (including social networking sites) and any other electronic storage devices;

d. Defendants are restrained from inducing or attempting to induce, or from causing any person or other entity to induce or attempt to induce, any person who is an employee
of Lost Creek to breach a contract with Lost Creek and to leave the employ of Lost Creek;
e. Weird is restrained from the design, development, testing, manufacture, promotion lease or sale of any products that resemble or imitate any pipeline pig manufactured, sold or developed by Lost Creek or providing the same or similar functions for Artemis that he performed for Lost Creek until March 1, 2015;
f. Defendants are ordered to cease and desist from leasing, selling, promoting, or otherwise commercially using the AirHog™ Remote-Controlled Pipeline Smart Pig or any other tool designed or derived by using Lost Creek’s trade secrets or confidential information, including but not limited to the WPF Proprietary Technology.

30. Upon information and belief, Defendants used, misappropriated, and disclosed Lost Creek’s trade secrets and/or proprietary confidential information and continue to do so for the purposes of furthering Artemis’ business. Defendants have solicited and continue to solicit Lost Creek’s customers. It is believed that Defendants may continue to solicit Lost Creek’s employees to breach contracts with Lost Creek in order to work for Artemis. The evidence of Defendants’ breach of contract, tortious interference, unfair competition, and/or misappropriation of trade secret claims support this Court’s granting of its request for injunction. Lost Creek would similarly be entitled to the requested relief after a trial on the merits.

31. If Lost Creek’s Application is not granted, harm is imminent because upon information and belief, Defendants are presently in possession of Lost Creek’s trade secrets, proprietary confidential information and/or have transmitted Lost Creek’s trade secrets, proprietary confidential information to others to facilitate their use of that information for their own benefit. In addition, upon information and belief, Defendants have solicited and continue to solicit Lost Creek’s former, current, and/or prospective customers and its employees. These actions are tortious and violate Weird’s fiduciary duties and/or contractual obligations to Lost Creek.

32. The harm that will result if the Permanent Injunction is not issued is in part irreparable. Lost Creek cannot be fully compensated for all such harm. Money cannot fully compensate Lost Creek for the loss of its trade secrets and proprietary confidential information, which Lost Creek invests substantial time, money, and human capital resources to develop, and which gives Lost Creek a competitive advantage in the marketplace and which, if used, gives to Defendants a commercial advantage. Lost Creek also cannot be fully compensated for the continued loss of its employees to Artemis. Lost Creek cannot be fully compensated by the loss of its goodwill that will result from the loss of its trade secrets, proprietary confidential information, employees, and business opportunities.
33. The injury Lost Creek faces outweighs the injury that would be sustained by the Defendants as a result of the injunctive relief. The injunctive relief sought would not adversely affect public policy or the public interest.

34. Lost Creek is willing to post the necessary reasonable bond to facilitate the above injunctive relief requested.

V. CAUSES OF ACTION

Count 1 - Breaches of Trade Secret Agreement and Covenant Not to Compete

35. The foregoing paragraphs are incorporated by reference as if fully stated herein.

36. The Non-Disclosure Agreement and Covenant Not to Compete executed and agreed to by Weird precludes Weird from competing against Lost Creek for a period of two (2) years. The Non-Disclosure Agreement and Covenant Not to Compete executed by Weird also include Weird’s promises not to disclose or use Lost Creek’s confidential information and trade secrets.

37. Weird’s Non-Disclosure Agreement and Covenant Not to Compete agreement is enforceable under Texas law. Weird’s promises in the agreement were each made in exchange for Lost Creek’s promises to provide Weird with specialized knowledge and training, Lost Creek’s trade secrets, Lost Creek’s proprietary confidential information and Lost Creek’s goodwill. Lost Creek fulfilled each of these promises with respect to Weird. Each of the covenants arise out of the trade secret agreement because the covenant is: (1) designed to protect Lost Creek’s trade secrets, Lost Creek’s confidential and proprietary information, Lost Creek’s goodwill, and the specialized training and knowledge Lost Creek provided to Weird; and (2) to enforce Weird’s promises regarding the same.

38. Weird’s covenants not to compete have reasonable time, territory, and activity limitations. The covenants’ limitations do not impose greater restraint than necessary to protect Lost Creek’s business interests; and Lost Creek does not seek to enforce the covenants in any unreasonable manner or to any unreasonable extent.

39. Upon information and belief, Weird violated his Non-Disclosure Agreement and Covenant Not to Compete by divulging, disclosing, and using trade secrets and/or proprietary confidential information as discussed above.

40. The above breaches are material. As a natural, probable, and foreseeable consequence and proximate cause of Weird’s actions, Lost Creek has suffered and continues to suffer damages for which Weird and Artemis are liable. Lost Creek seeks to recover all special, general, consequential, actual, and exemplary damages allowed by law as well as attorney fees, court costs, prejudgment, and post-judgment interest. Lost Creek has or will suffer damages to its
business in the form of lost profits, loss of customers, loss of future business opportunities, loss of the exclusive right to use Lost Creek’s trade secrets, and loss of goodwill. Lost Creek seeks to recover lost profits from contracts that were awarded to Artemis as a result of Weird’s breaches of contract. In order to fully develop its lost profit claims, Lost Creek must examine Artemis’ documents to determine the value of the jobs Artemis obtained. In the alternative, and in the event that Lost Creek’s lost profits are unascertainable, Lost Creek seeks unjust enrichment damages.

**Count 2 – Unfair Competition by Misappropriation**

41. The foregoing paragraphs are incorporated by reference as if fully stated herein.

42. An employee's employment relationship with his or her employer gives rise to a duty that forbids an employee from using his employer's trade secrets or any other confidential or proprietary information of the employer acquired during the employment relationship in competition with the employer or in any other manner adverse to the employer. This common law duty survives the termination of employment.

43. As alleged above, Defendant Weird has engaged in unfair competition through his knowing and intentional breaches of these common-law duties. Plaintiffs have been damaged in an amount that exceeds the minimum jurisdictional limits of this Court and are entitled to a permanent injunction as requested.

**Count 3 – Tortious Conversion**

44. The foregoing paragraphs are incorporated by reference as if fully stated herein.

45. As alleged above, Plaintiff owned trade secrets and other confidential and proprietary information. Defendants assumed and exercised dominion and control over Plaintiffs trade secrets and other confidential information in an unlawful and unauthorized manner. Plaintiff has been damaged in an amount that exceeds the minimum jurisdictional limits of this Court.

**Count 4 - Common Law Misappropriation of Trade Secrets**

46. The foregoing paragraphs are incorporated by reference as if fully stated herein.

47. Lost Creek has suffered and continues to suffer damages that are a natural, probable, and foreseeable consequence and proximate cause of Defendants’ use and disclosure of Lost Creek’s trade secrets and confidential information. Lost Creek seeks to recover all special, general, consequential, actual, and exemplary damages allowed by law as well as attorney fees, court costs, prejudgment interest, and post-judgment interest. In particular, Lost Creek seeks damages based on the value of misappropriated trade secrets when they were
misappropriated; the diminution in the value of Lost Creek’s trade secrets to Lost Creek as a result of the misappropriation and disclosure by Defendants; the lost profits Lost Creek has suffered as a result of Defendants’ misappropriation, the disgorgement of Defendants’ profits associated with the use of Lost Creek’s trade secrets, a reasonable royalty which Defendants would have been willing to pay and Lost Creek would have been willing to accept for the use of Lost Creek’s trade secrets; and Defendants’ “unjust enrichment” resulting from the misappropriation of Lost Creek’s trade secrets. Unjust enrichment includes the following: (1) Defendants’ profits resulting from the use of the trade secrets; (2) Defendants’ profits on sales made possible by product development which was accelerated by the misappropriation of the trade secrets; and/or (3) avoided development costs resulting from the misappropriation.

48. In addition to these damages, Lost Creek seeks permanent injunctive relief to prevent all such imminent and irreparable harm in the future.

Count 5 - Tortious Interference with M-I’s Employment Contracts

49. The foregoing paragraphs are incorporated by reference as if fully stated herein.

50. Lost Creek had valid contracts with the aforementioned employees, including but not limited to its Non-Disclosure Agreement and Covenant Not to Compete agreements and/or at will employment agreements. Artemis and its agents, including Montgomery Bonnell, knew or had reason to know of the above contracts, specifically the Non-Disclosure Agreement and Covenant Not to Compete, because Bonnell obtained the agreement from Weird when Weird was hired and while Bonnell was an employee of Lost Creek. Further, Bonnell executed essentially the same agreement with Lost Creek when he was employed by Lost Creek. Artemis and its agents willfully and intentionally interfered with the contracts. Artemis and its agents induced the former employees to quit Lost Creek and join Artemis. Artemis offered them increased compensation and/or other benefits. The former employees perform or performed the same duties for Artemis they did for Lost Creek. These former employees are violating or have violated their covenants not to compete. Upon information and belief, the former Lost Creek employees have used and continue to use Lost Creek’s confidential information and trade secrets in their employment with Artemis.

Count 6 – Breach of Fiduciary Duty

51. The foregoing paragraphs are incorporated by reference as if fully stated herein.

52. Weird and Bonnell, agents of Artemis and former employees of Lost Creek, each owed Lost Creek a fiduciary duty. This fiduciary duty survives termination of employment with Lost Creek. This fiduciary duty includes, among other things, a duty not to: (1)
Lost Creek’s trade secrets and confidential information; (2) solicit the departure of other Lost Creek employees while working for Lost Creek; or (3) form a competing enterprise.

53. Upon information and belief, Weird, Bonnell and agents of Artemis breached their respective fiduciary duties to their benefit by appropriating Lost Creek’s trade secrets and confidential information and soliciting or obtaining the departure of other Lost Creek employees. Further, weird breached his fiduciary duty to Lost Creek by fostering a competing enterprise while employed with Lost Creek.

**Count 7 – Civil Conspiracy**

54. The foregoing paragraphs are incorporated by reference as if fully stated herein.

55. Defendants have secretly and intentionally conspired, agreed, and endeavored to interfere with Lost Creek’s prospective business relationships and contracts and employee contracts, deprive Lost Creek of business goodwill, and damage Lost Creek’s reputation. This conspiracy has proximately caused Lost Creek to suffer damages.

56. Defendants, agreed to interfere with Lost Creek’s prospective contracts with Lost Creek’s customers and Lost Creek’s contracts with its employees. Defendants knew that this interference would result in harm to Lost Creek. Lost Creek has suffered, and continues to suffer, damages that are proximately caused by Defendants’ conspiracy to interfere with Lost Creek’s contracts with its current, former, and prospective customers and employees. Lost Creek seeks to recover all special, general, consequential, actual, and exemplary damages allowed by law as well as court costs, prejudgment interest, and post judgment interest. Lost Creek has or will suffer an amount of damages to its business in the form of lost profits, loss of customers, loss of future business opportunities, loss of the exclusive right to use its trade secrets, and loss of goodwill.

**Count 9 - The Computer Fraud and Abuse Act - 18 U.S.C. § 1030**

57. The foregoing paragraphs are incorporated by reference as if fully stated herein.

58. Lost Creek’s computers are used in interstate commerce; thus, Lost Creek’s computers are protected computers pursuant to 18 U.S.C. § 1030 (e)(2)(B).

59. Weird knowingly and with intent to defraud, accessed and used the computer(s) assigned by Lost Creek, without authorization or in a manner exceeding any authorization he may claim that he had. By means of such conduct, Weird furthered the intended fraud.
60. Lost Creek believes that, in August 2015 and on other occasions, Weird used Lost Creek’s computer(s) to misappropriate, use, and share Lost Creek’s trade secrets and proprietary confidential information without authorization.

61. Because of Weird’s actions, Lost Creek suffered losses in excess of $5,000, including costs related to a computer forensic preservation and analysis of Weird’s Lost Creek issued laptop and iPhone.

VI. ATTORNEY FEES AND INTEREST

62. Pursuant to statute, common law, and the contracts with Defendants, Plaintiff is entitled to an award of its reasonable and necessary attorney fees with respect to Defendants for this cause and any appeals.

VII. EXEMPLARY DAMAGES

63. The conduct of Defendants, as alleged above, including tortious interference with employee contracts, tortious interference with prospective business relationships and contracts, misappropriation and disclosure of trade secrets, and civil conspiracy, was aggravated by the kind of willfulness, wantonness and malice for which the law allows for the imposition of exemplary damages. Moreover, Defendants’ wrongdoing was committed knowingly and with a conscious indifference to Lost Creek’s rights. Defendants acted with intent to harm Lost Creek and their misconduct and tortious interference was intentional, willful, wanton and without justification or excuse. Therefore, Lost Creek seeks to recover exemplary damages from Defendants in an amount to be determined by the Court.

VIII. CONDITIONS PRECEDENT

64. All conditions precedent to an outcome favorable to the party represented by the undersigned in this action have been performed, have occurred or have been waived.

IX. PRAYER

WHEREFORE, PREMISES CONSIDERED, Plaintiff prays for the following relief:

a) A permanent injunction for the relief requested above;

b) Upon final trial, judgment against Defendants, jointly and severally, for full permanent injunctive relief as requested herein, and, for the full amount of the Plaintiff’s damages, special, general, consequential, actual, and exemplary;

c) Assignment of Defendants' Provisional Patent Application and/or Patent on the AirHog™ Remote-Controlled Pipeline Smart Pig and/or related technologies;

d) Prejudgment interest;

e) Post judgment interest;
f) Plaintiff’s reasonable and necessary attorney fees in prosecuting its claims through trial and, if necessary appeal;
g) All costs of suit; and
h) Such other and further relief, at law or in equity, to which Plaintiff may show itself justly entitled.

Respectfully submitted,
Massive International Law Firm, LLP

/S/
By: _________________________
Ben Dover
TSB No. 00003723
1 Congress Ave., Suite 20000
Austin, Texas Austin 78701
Tel: (512) 555-1234
bdover@milf.com
LEAD ATTORNEYS FOR PLAINTIFF

Of Counsel
Lamar Street
Tarrytown, Olde & Rich, Attorneys
1313 Guadalupe, Suite 1900
Austin, Texas 78701

CERTIFICATE OF SERVICE
I certify that a true copy of the foregoing has been served on all attorneys of record and persons pro se in this cause, by electronic service, electronic mail, facsimile and/or certified mail, return receipt requested, by depositing same, postpaid, in an official depository under the care and custody of the United States Postal Service on March 15, 2016.

//s// Ben Dover
IN THE UNITED STATES DISTRICT COURT
WESTERN DISTRICT OF TEXAS

LOST CREEK ENGINEERING, L.L.C. §

§

V. § CIVIL ACTION NO.

§ 5:16-CV-00001

KEITH AUSTIN WEIRD, §

and § JURY REQUESTED

ARETEMIS ENERGY SOLUTIONS, INC. §

§

FIRST AMENDED ANSWER OF KEITH AUSTIN WEIRD AND ARTEMIS ENERGY SOLUTIONS, INC.

Defendants Keith Austin Weird and Artemis Energy Solutions, Inc., file this Amended Answer in response to the Amended Complaint and Application for Injunctive Relief filed by Plaintiff, Lost Creek Engineering, Inc. ("Lost Creek").

I. FIRST AMENDED ANSWER

1. Defendants are not required to admit or deny the allegations of Paragraph 1.
2. Defendants admit the allegations of paragraph 2.
3. Defendants admit the allegations of paragraph 3
4. Paragraph 4 is a statement of jurisdiction which Defendants are not required to admit or deny.
5. Paragraph 5 is a statement of venue which Defendants are not required to admit or deny.
6. Upon information and belief, Defendants admit the allegations of Paragraph 6.
7. Upon information and belief, Defendants admit the allegations of Paragraph 7.
8. With regard to Paragraph 8, Defendants deny the allegation that Lost Creek’s When Pigs Fly (WPF) smart pig technology, if any, represent unique or innovative technology. Defendants admits all other allegations in Paragraph 8.

426
9. With regard to paragraph 9, Defendants deny that Lost Creek’s WPF technologies (if any) are not found in the open market and are not used by other smart pig manufacturers. Defendants contend that all or part of these allegedly proprietary and confidential WPF technologies (if any) derive from open sources and/or were not developed by Lost Creek. Defendants admit all other allegations in Paragraph 9.

10. Defendants deny all allegations in Paragraph 10.

11. Defendants deny that the information referred to as “e WPF Proprietary Technology “is not known outside of the Lost Creek and provides a competitive advantage to Lost Creek in the marketplace. Defendants do not have sufficient information to either admit or deny the other allegations in Paragraph 11.

12. Defendants deny that Lost Creek has taken great care to ensure that the custom design features of its products and manufacturing processes are kept confidential and remain a trade secret. Defendants deny that Lost Creek’s WPF Proprietary Technology and other details of Lost Creek’s custom features cannot be found in the open market and are not available to competitors to view or reverse engineer. Defendants admit all other allegations in Paragraph 12.

13. With regard to Paragraph 13, Defendants admit that Weird executed a Non-Disclosure Agreement and Covenant Not to Compete on September 1, 2000 after being ordered to do so by Lost Creek. Defendants admit that Non-Disclosure Agreement and Covenant Not to Compete is an industry-wide, unenforceable restraint of trade that purports to forbid Weird from competing directly or indirectly with Lost Creek for a period of two years, without territorial restriction. Defendants do not have sufficient information to either admit or deny any other allegations in Paragraph 13.


15. Defendants admit the allegations of paragraph 15.

16. Defendants deny that the information referenced as “unique WPF Proprietary Technology’ is unique, proprietary or the property of Plaintiff Lost Creek. Defendants admit the other allegations of paragraph 16.

17. Defendants admit the allegations of paragraph 17.
18. Defendants deny that Weird downloaded almost thirty gigabytes of data comprising thousands of Lost Creek’s confidential business documents and trade secrets. Defendant admits that he may have sought to back up certain iTunes music he personally purchased as well as family photographs. Defendants contend that any business documents copied by Weird were either copied inadvertently or were copied for the purpose of completing work for Lost Creek’s sole and exclusive benefit. Defendants do not have sufficient information to either admit or deny any other allegations in Paragraph 18.

19. Defendants admit the allegations of paragraph 19.

20. Defendants deny all allegations of paragraph 20.

21. Defendants do not have sufficient information to either admit or deny the allegations in Paragraph 21. Defendants admit that over the course of 15 years of employment, Weird may have used his personal e-mail for his former employer’s benefit.

22. Defendants deny that there have been repeated requests made for the return of Weird’s Lost Creek laptop or that Weird has declined (or failed to respond to) requests for inspection. Many of the devices and sources described hold confidential personal and privileged information and communications. Defendants admit that Weird participated in an exit interview.

23. Defendants admit the allegations of paragraph 23.

24. Defendants deny all allegations of paragraph 24.

25. Defendants deny all allegations of paragraph 25.

26. Defendants deny that prior to Weird’s employment with Artemis, Artemis did not manufacture or sell any type of smart pig products that competed with Lost Creek’s products. Defendants admit all other allegations of Paragraph 26.

27. Defendants admit Artemis has applied for a patent on unique and innovative design features of certain of its intelligent pipeline pig products. Defendants deny all other allegations of Paragraph 27.

28. Defendants incorporate their prior responses to Paragraphs 1-27.

29. Defendants deny all allegations of paragraph 29.

30. Defendants deny all allegations of paragraph 30.

31. Defendants deny all allegations of paragraph 31.
32. Defendants deny all allegations of paragraph 32.
33. Defendants deny all allegations of paragraph 33.
34. Defendants do not have sufficient information to either admit or deny any allegations in Paragraph 34.
35. Defendants incorporate their prior responses to Paragraphs 1-34.
36. Defendants admit the allegations of paragraph 36.
37. Defendants deny all allegations of paragraph 37.
38. Defendants deny all allegations of paragraph 38.
39. Defendants deny all allegations of paragraph 39.
40. Defendants deny all allegations of paragraph 40.
41. Defendants incorporate their prior responses to Paragraphs 1-40.
42. Defendants do not have sufficient information to either admit or deny any allegations in Paragraph 42.
43. Defendants deny all allegations of paragraph 43.
44. Defendants incorporate their prior responses to Paragraphs 1-43.
45. Defendants deny all allegations of paragraph 45.
46. Defendants incorporate their prior responses to Paragraphs 1-45.
47. Defendants deny all allegations of paragraph 47.
48. Defendants deny all allegations of paragraph 48.
49. Defendants incorporate their prior responses to Paragraphs 1-48.
50. Defendants deny all allegations of paragraph 50.
51. Defendants incorporate their prior responses to Paragraphs 1-50.
52. Defendants deny all allegations of paragraph 52.
53. Defendants deny all allegations of paragraph 53.
54. Defendants incorporate their prior responses to Paragraphs 1-53.
55. Defendants deny all allegations of paragraph 55.
56. Defendants deny all allegations of paragraph 56.
57. Defendants incorporate their prior responses to Paragraphs 1-56.
58. Defendants deny all allegations of paragraph 58.
59. Defendants deny all allegations of paragraph 59.
60. Defendants deny all allegations of paragraph 60.
61. Defendants deny all allegations of paragraph 61.
62. Defendants deny all allegations of paragraph 62.
63. Defendants deny all allegations of paragraph 63.

II. AFFIRMATIVE DEFENSES

FIRST AFFIRMATIVE DEFENSE: FAILURE TO STATE A CLAIM.
64. Defendants affirmatively assert that Lost Creek’s claims are barred, in whole or in part, because Lost Creek has failed to state a claim upon which relief may be granted.

SECOND AFFIRMATIVE DEFENSE: WAIVER.
65. Defendants affirmatively asserts that Lost Creek’s claims are barred by the doctrine of waiver.

THIRD AFFIRMATIVE DEFENSE: ESTOPPEL.
66. Defendants affirmatively assert that Lost Creek’s claims are barred by the doctrine of estoppel.

FOURTH AFFIRMATIVE DEFENSE: INJUNCTIVE RELIEF IS UNECESSARY.
67. Defendants affirmatively assert that Lost Creek’s claims are barred, in whole or in part, because injunctive relief is unnecessary as pled.

FIFTH AFFIRMATIVE DEFENSE: JUSTIFICATION.
68. Defendants affirmatively asserts that Lost Creek’s claims are barred because of the doctrine of justification.

SIXTH AFFIRMATIVE DEFENSE: PRIVILEGE.
69. Defendants affirmatively assert that Lost Creek's claims are barred, in whole or in part, because of the doctrine of privilege.

SEVENTH AFFIRMATIVE DEFENSE:

PREEMPTION OF ATTORNEY’S FEES AWARD.
70. Defendants affirmatively assert that Lost Creek’s claims for attorney's fees are barred, in whole or in part, because such claims are preempted by the Texas Covenant not to Compete Act. TEX. Bus. & COM. CODE ANN. § 15.51 and § 15.52.
III. ATTORNEY'S FEES

71. The primary purpose of the "agreement" to which Lost Creek claims the Non-Disclosure Agreement and Covenant Not to Compete was ancillary to, was to obligate Weird to render personal services. Plaintiff knew that the Non-Disclosure Agreement and Covenant Not to Compete did not contain limitations as to time, geographical area, and scope of activity to be restrained that were reasonable and the limitations imposed a greater restraint than necessary to protect the goodwill or other business interest of Plaintiff.

72. Plaintiff is also seeking to enforce the covenant to a greater extent than is necessary to protect Plaintiff's goodwill or other business interest. Therefore, Pursuant to Section 15.51 of the Texas Business and Commerce Code, Defendants seek to recover reasonable attorney's fees and costs as are equitable and just.

73. Additionally, Defendants seek to recover reasonable attorney's fees and costs pursuant to Section 134.005 of the Texas Civil Practice & Remedies Code.

Respectfully Submitted,

Bevo ★ Orange ★ Tower, P.C.

By: [Signature]
"Tex" S. Tower
Federal ID No. 123456
State Bar No. 010101010
2300 Inner Campus Drive
Austin, Texas 78713
TEL: (512) 555-3377

ATTORNEY IN CHARGE FOR DEFENDANTS KEITH AUSTIN WEIRD AND ARTEMIS ENERGY SOLUTIONS, INC.

CERTIFICATE OF SERVICE

The undersigned hereby certifies that a true and correct copy of the above and foregoing was served pursuant to the Federal Rules of Civil Procedure on this the 18th day of March, 2016, to:

Ben Dover
1 Congress Ave., Suite 20000
Austin, Texas Austin 78701

[Signature]