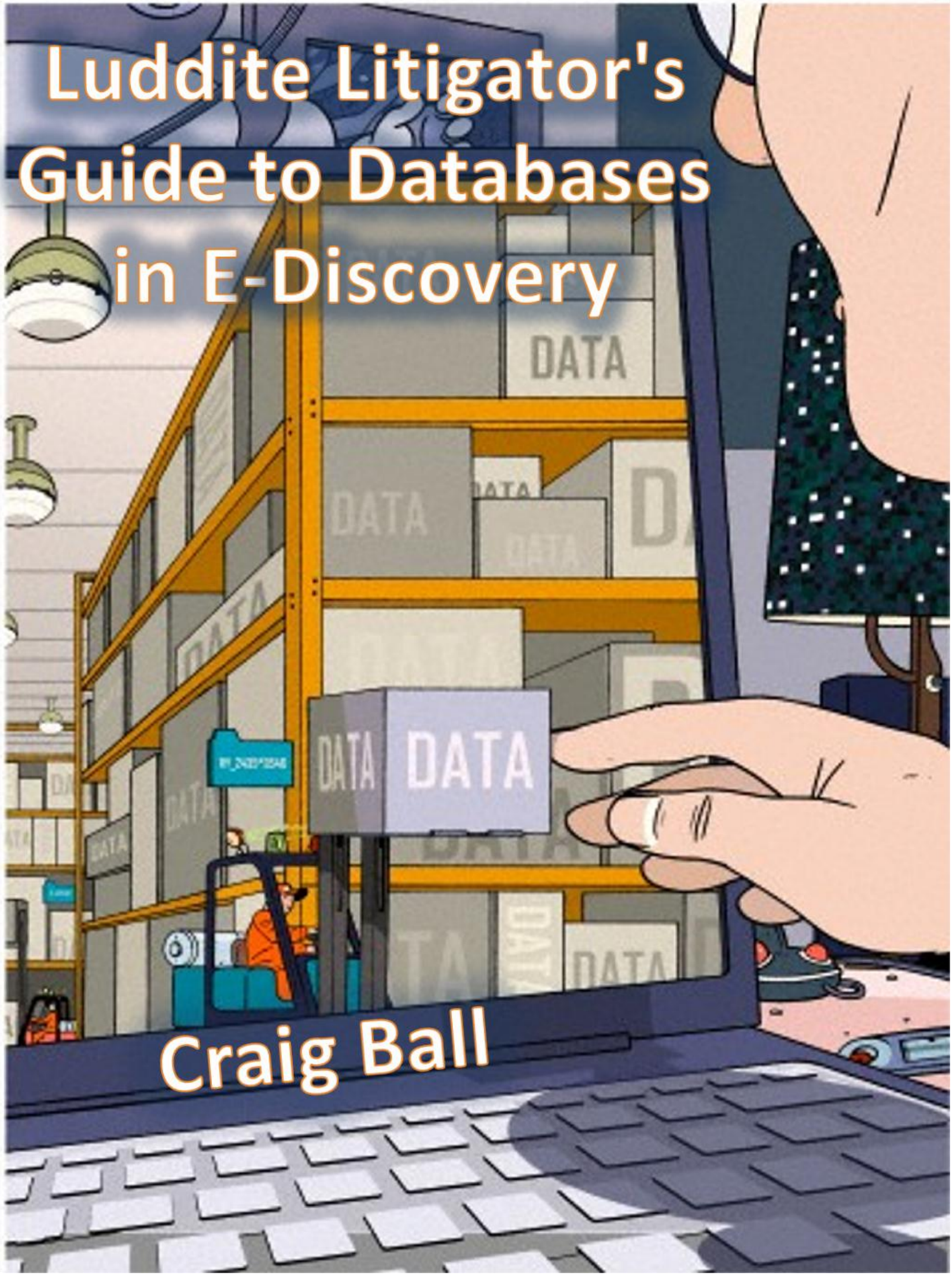
Luddite Litigator's Guide to Databases in E-Discovery



Craig Ball

The Luddite Litigator's Guide to Databases in E-Discovery

By Craig Ball

When I set out to write a paper 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.[1] A few commentators mention *In re Ford Motor Co.,*[2] 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

[1] A noteworthy exception is the discussion of databases in Michael Arkfeld's superb treatise, *Arkfeld on Electronic Discovery and Evidence*, §3.11 (2d Ed.).

[2] 345 F.3d 1315 (11th Cir. 2003)

information assembled 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.[3]

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:

CLIENTS

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

It's essential to keep track of cases and upcoming trials, so you create another table called CASES:

[3] 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.

CASES

CASE_NO	TRL_DATE	MATTER	TYPE	COURT
001	2011-02-14	U.S. v. Microsoft	Antitrust	FDDC-1
002	2012-01-09	EON v Cisco	Patent	FEDTX-2
003	2011-02-15	In re: Dell	Regulatory	FWDTX-4
004	2011-05-16	SAP v. Oracle	Conspiracy	FNDCA-8
005	2012-01-09	Microsoft v. Yahoo	Breach of K	FWDWA-6
006	2010-12-06	Apple v. Adobe	Antitrust	FNDCA-8
007	2011-10-31	Dell v. Travis County	Tax	TX250
008	null	Hawkins v. McGee	Med Mal	FUSSC
009	2011-12-05	Jobs v. City of Woodside	Tax	CASMD09

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:

COURTS

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

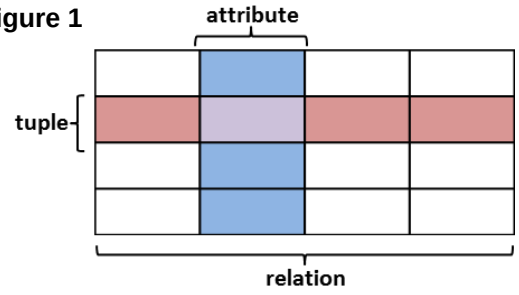
As we look at these three tables, note that each has a unique key field called the "**primary key**" for that table.[4] For the CLIENTS table, the primary key is the client's last name.[5] The primary key is the trial date for the TRIAL_DATES table and it's a unique court identifier for the COURTS table. The essential characteristic of a primary key is that it cannot repeat within the table for 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.

[4] Tables can have more than one primary key.
[5] 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.

Tuples and Attributes

Figure 1

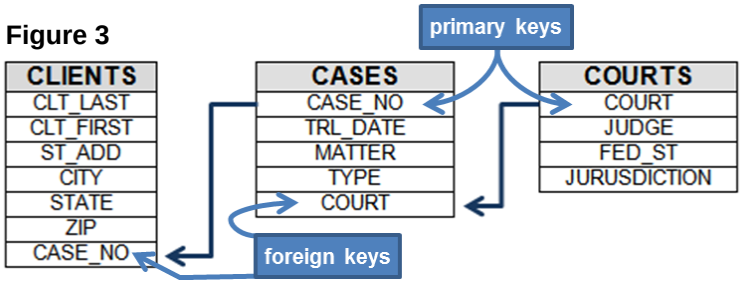


Now, we need to introduce some new terminology because the world of relational databases has a language all its own. Dealing with the most peculiar term first, the contents of each row in a table is called a **"tuple,"** defined as an ordered list of elements.[6] In the COURTS table above, there are 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).

Figure 2

The magic happens in a relational database when tables are **"joined"** (much like the cube in Figure 2)[7] by referencing one table from another.[8] This is done by incorporating the primary key in the table referenced as a **"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

Figure 3



---

[6] 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 we went to law school instead of studying computer science.

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

[8] The term "relation" is so confounding here, I will continue to refer to them as tables.

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 serves 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).

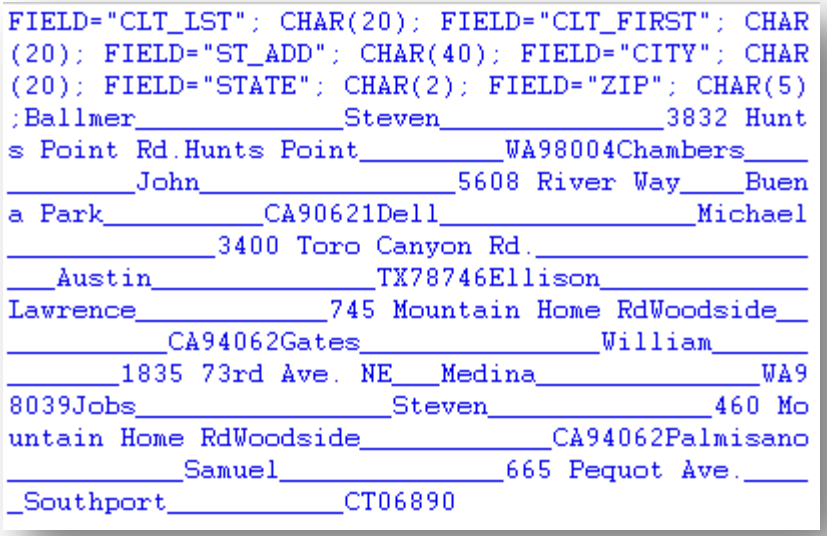
You'll appreciate why demanding production of the raw tables in a database may be an untenable approach to e-discovery when you consider how databases store information. When a database populates a table, it's stored in either **fixed length** or **variable length** fields.

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).

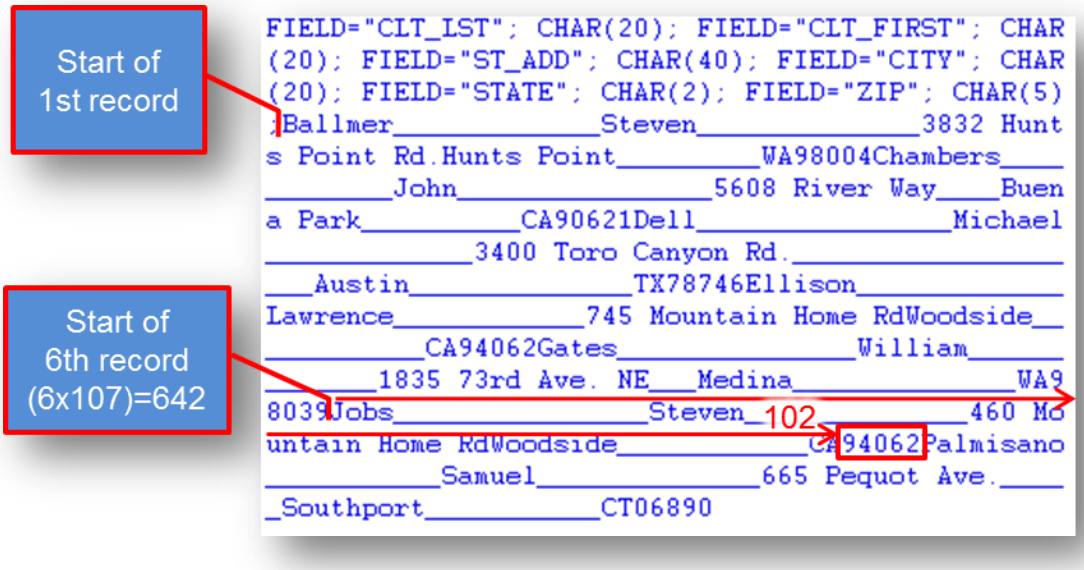
A fixed length field record for a simplified address table might look like Figure 4.
Note how the data is one continuous stream. The name, order and length of data allocated for each field is defined at the beginning of the string in all those "FIELD=" and CHAR(x) statements, such that the total length of each record is 107 characters. To find a given record in a table, the database software simply starts accessing data

Figure 4



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:

```
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.

.

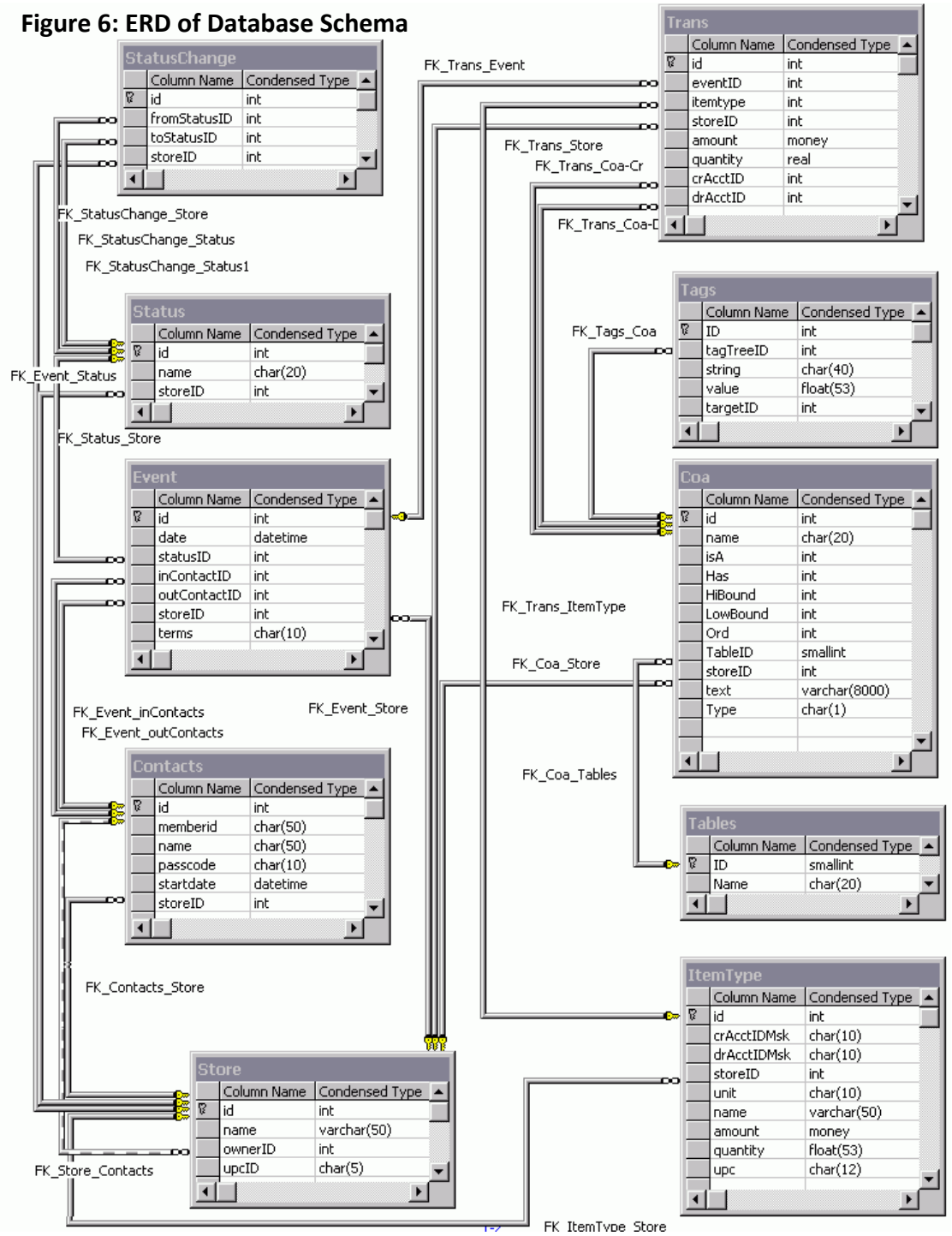
In the lingo of database land, the “map” is the database’s **schema,** and it’s housed in the system’s **data dictionary**. It may be the system’s **logical schema**, detailing how the database is designed in terms of its table structures, attributes, fields, relationships, joins and views. Or, it could 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 InfrastructureasaService (“IaaS”) 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 courts 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.