Richmond Journal of Law and Technology

Volume 19 | Issue 3 Article 2

2013

Databases Lie! Successfully Managing Structured Data, the Oft-Overlooked ESI

Conrad Jacoby

Jim Vint

Michael Simon

DATABASES LIE! SUCCESSFULLY MANAGING STRUCTURED DATA, THE OFT-OVERLOOKED ESI

By Conrad Jacoby,* Jim Vint,** & Michael Simon***

Cite as: Conrad Jacoby, Jim Vint & Michael Simon, *Databases Lie! Successfully Managing Structured Data, The Oft-Overlooked ESI*, 19 RICH. J.L. & TECH 9 (2013), *available at* http://jolt.richmond.edu/v19i3/article9.pdf.

[1] Legal professionals regularly advise clients to ensure that the storage, retention, and accessibility of their Electronically-Stored Information ("ESI") is in full compliance with all legal and regulatory requirements in the event this information becomes relevant in civil, criminal, or regulatory disputes. However, what many practitioners may not realize is that the ESI that clients are required to produce for e-discovery includes both "unstructured" and "structured" data. Searching and producing only one of these types of ESI may well not fully satisfy a client's full discovery obligations. Even worse, it might not present a full understanding of the factual issues in the matter and how to best prove them to the legal team.

I. WHAT IS "STRUCTURED DATA?"

[2] Most legal professionals are extremely familiar with "unstructured" or "loose" data, even if they do not necessarily know it by these terms. Simply put, unstructured data refers to e-mail messages, word processing documents, spreadsheets, and presentations, among other things—in other words, human-readable information that is commonly sought as potentially relevant ESI in discovery.[1] Structured data, on the

* Conrad Jacoby is a Senior Attorney at Winston & Strawn LLP, where his practice focuses on e-discovery issues and litigation information management. Since 2009, he has served as the founding Editor-in-Chief of *The Sedona Conference® Database Principles: Addressing the Preservation & Production of Databases & Database Information in Civil Litigation*. The opinions expressed are those of the authors and do not necessarily

other hand, refers to information residing in electronic repositories or silos, such as transactional and financial databases.[2] Unlike unstructured data, which typically exists as static and self-contained files that are preserved, collected, processed, reviewed, authenticated, and admitted into evidence as stand alone documents, structured data exists as segments of information inside a larger system, one that is often quite complex and contains many parts.[3] A database record, the closest analog that structured data has to a "document," may not actually exist until a user performs some action through the database system to assemble a number of separate fields that could reside in many different parts of the system. For this reason, information stored in a database cannot be placed into a standard e-discovery review system that has been optimized to view and categorize unstructured data.

---

reflect the views of their respective firms or clients. This article is for general information purposes and is not intended to be and should not be taken as legal advice

** Jim Vint is a Managing Director at Navigant Consulting, Inc. and runs the Structured Data and Development team within the Technology Solutions group. He focuses on discovery and disclosure of non-traditional ESI data sources including structured databases. His clients include global organizations facing regulatory investigations, cross border discovery issues, and general commercial disputes.

*** Michael Simon is Director of Strategic Development for Navigant Consulting, Inc. Michael, a former practicing attorney, has worked with and counseled clients regarding e-discovery issues and best practices for over a decade. He frequently lectures on e-discovery, legal technology and Internet law in venues across the United States, including Tufts University, where he has taught as a visiting lecturer.

[1] *See* THE SEDONA CONFERENCE®, THE SEDONA CONFERENCE® GLOSSARY: E-DISCOVERY AND INFORMATION MANAGEMENT 52 (Sherry B. Harris ed., 3d ed. 2010) [hereinafter *Sedona Glossary*].

[2] *See id.* at 49.

[3] *See id.* at 13, 49, 52 (definitions of "database," "database management system," "structured data," and "unstructured data").

[3] The ESI stored in databases and other structured data repositories is every bit as relevant and discoverable as the loose files that are more commonly requested. Federal Rule of Civil Procedure ("FRCP") 34 is clear and unambiguous on this point:

> Rule 34. Producing Documents, Electronically Stored Information, and Tangible Things, or Entering onto Land, for Inspection and Other Purposes
>
> (a) In General. A party may serve on any other party a request within the scope of Rule 26(b):
>
> (1) to produce and permit the requesting party or its representative to inspect, copy, test, or sample the following items in the responding party's possession, custody, or control:
>
> (A) any designated documents or electronically stored information—including writings, drawings, graphs, charts, photographs, sound recordings, images, and other **data or data compilations**—stored in any medium from which information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably usable form[.][4]

[4] Unlike the discovery of unstructured data, for which a number of best practices have emerged, it has been difficult for the legal industry to develop best practices for the treatment of structured data in civil discovery due to the vast diversity of size, scope, and features found in different database systems. The Sedona Conference®, a non-partisan legal think-tank founded in 1997, formed a group in early 2009 to study the issues surrounding the discovery of structured data—culminating in the publication of *The Sedona Conference® Database Principles Addressing the Preservation and Production of Databases and Database Information in Civil Litigation* (hereinafter the "*Sedona Database Principles*") in April

[4] FED. R. CIV. P. 34(a)(1)(A) (emphasis added).

2011.[5] The *Sedona Database Principles* expand upon the original publication, *The Sedona Principles: Best Practices Recommendations & Principles for Addressing Electronic Document Production* (hereinafter the "*Sedona Principles*"),[6] as they specifically apply to databases and set out six additional precepts that provide practical suggestions for simplifying the discovery of structured data and clarifying the obligations of both the requesting and producing parties.[7] An overarching theme of the *Sedona Database Principles* is that better communication between parties, their legal advisors and agents, and information technology professionals will substantially improve the management of this type of specialized ESI in legal disputes.[8] To that end, the *Sedona Database Principles* specifically reference many of the precepts of the *Sedona Principles* that address and encourage cooperation between the parties.[9]

II. HOW DOES STRUCTURED DATA BECOME RELEVANT?

[5] Databases frequently record historical transactions and information that is relevant in litigation and investigations. One would certainly expect that enterprise-level systems like Oracle and SAP, not to mention financial and transactional systems, human resource tracking systems, data warehouses, and content management systems ("CRM"), would all contain structured data. However, other commonly used systems,

[5] *See* THE SEDONA CONFERENCE®, THE SEDONA CONFERENCE® DATABASE PRINCIPLES: ADDRESSING THE PRESERVATION & PRODUCTION OF DATABASES & DATABASE INFORMATION IN CIVIL LITIGATION 21 (Conrad J. Jacoby et al. eds., 2011) [hereinafter *Sedona Database Principles*].

[6] *See* THE SEDONA CONFERENCE®, THE SEDONA PRINCIPLES: BEST PRACTICES RECOMMENDATIONS & PRINCIPLES FOR ADDRESSING ELECTRONIC DOCUMENT PRODUCTION 30 (Jonathan M. Redgrave et. al ed., 2d eds. 2007) [hereinafter *Sedona Principles*].

[7] *See Sedona Database Principles*, *supra* note 5, at ii, 8.

[8] *See id.* at ii.

[9] *See id* at ii, 8-9.

including Cloud-based "Software-As-A-Service" ("SaaS") systems, also feature the same back-end structured data systems as more obvious "database" systems. Thus, structured data has largely replaced loose documents for tracking information for these and other similar functions: accident/incident reporting systems, call center records and associated data analytics, world wide web servers, point of sale systems, and social media.

[6] The cumulative volume of data in business-related structured data repositories is immense and is projected to grow at an estimated annual rate of nearly twenty percent.[10] Perhaps even more important to e-discovery practitioners, a recent survey about the state of discovery in civil litigation has shown that e-mail, the central focus of e-discovery requests for over fifteen years, is no longer the leading requested item.[11] Instead, database and application data are now more often requested.[12]

[7] An increasing number of litigation disputes involving "high profile" companies have made demands upon litigants to review, disclose, and produce at least portions of their databases. Several examples are explored below.

[8] The plaintiffs in *In re eBay Seller Antitrust Litigation*, an antitrust class action, sought production of transactional data from defendant eBay.[13] The court granted the motion in part and eBay objected, claiming that the information sought did not already exist in easily compiled form,

[10] Nexsan Corp., Registration Statement (Form S-1), at 61 (Jan. 25, 2011), *available at* http://www.sec.gov/Archives/edgar/data/1133448/000104746911000283/a2200385zex-99_2.htm.

[11] *See Information Retention and eDiscovery Survey Global Findings*, SYMANTEC 1, 8 (2011), https://www4.symantec.com/mktginfo/whitepaper/InfoRetention_eDiscovery_Survey_Report_cta54646.pdf.

[12] *Id.*

[13] *In re* eBay Seller Antitrust Litig., No. C 07-1882 JF (RS), 2009 WL 3613511, at *1 (N.D. Cal. Oct. 28, 2009).

requiring eBay "to spend hundreds of thousands of dollars to dedicate a highly specialized engineering resource for a period of more than six months to create new data" solely for the matter.[14] However, eBay's own submissions in support of the objection contained three different estimates, ranging from a low of $179,000 to a high of $300,000.[15] Moreover, eBay's employee in charge of data warehouse development declared that the provided estimate could vary "by as much as five hundred percent."[16] The court first disposed of eBay's argument that it could not be required to create anything new, finding that FRCP 34(a)(1)(A) supported the magistrate's finding that the technical burden of creating the new material did not excuse production.[17] In light of the hundreds of millions of dollars at stake in the action involving a defendant with billions of dollars in annual gross profits, and considering that the magistrate had already scaled back the scope of discovery, the court found no clear error in the magistrate's determination that the potential costs and technical requirements were not unduly burdensome.[18]

[9] In another case, a plaintiff injured by a sink that fell from a high storeroom shelf sought production of the database that the defendant, Lowe's, used to record and track accident and injury claims.[19] The trial court ordered Lowe's to present a witness with knowledge and access to the system and to print out all requests for accidents occurring before the date of the plaintiff's injury.[20] Notably, Lowe's objected that: (1) it had already produced a printout from the database of all falling merchandise

[14] *Id.*

[15] *Id.* at *2.

[16] *Id.*

[17] *See id.*

[18] *See In re eBay Seller Antitrust Litig.*, 2009 WL 3613511, at *3.

[19] *In re* Lowe's Cos., 134 S.W.3d 876, 877 (Tex. App. 2004).

[20] *See id.* at 877.

claims for its stores within the state for the last five years; (2) the remaining portions of the database were not relevant; (3) the manner in which accident information was gathered and stored was a trade secret; (4) the purpose of the database was not for safety-related information; and (5) there was no way to restrict production of privileged or non-relevant information.[21] The appellate court agreed with Lowe's in part and limited the plaintiffs from accessing data without limitation as to time, place, or subject matter.[22]

[10] In *Procter & Gamble v. Haugen* a plaintiff appealed from the dismissal of his Lanham Act and tortious interference claims which resulted in part from the court sanctioning it for failing to preserve relevant database information.[23] Procter & Gamble ("P&G") claimed that agents of a competitor spread false rumors that the company supported Satanism, using the profits from forty-three products to do so.[24] P&G and its expert witnesses used the services of a third party vendor, Information Resources Incorporated ("IRI"), to track potential lost sales of the forty-three involved products.[25] IRI used a database that gathered purchase information from retail stores into electronic market share databases.[26] IRI's databases stored data on a "rolling" basis so that data was kept only for a period of time before it was deleted from the system to make room for more data.[27] Defendants requested production of all of the information that P&G used from the IRI databases and when P&G was unable to produce all of this information, the court found that P&G had spoliated the

[21] *Id.* at 878.

[22] *See id.* at 880.

[23] Procter & Gamble Co. v. Haugen, 427 F.3d 727, 730, 732-37 (10th Cir. 2005).

[24] *Id.* at 731.

[25] *Id.* at 731-32.

[26] *Id*. at 731.

[27] *Id*.

data and dismissed the matter as a sanction.[28] On appeal, P&G focused on the fact that it was only a subscriber to the IRI database, did not own or control the system, and therefore could not have practicably provided the information to defendants.[29] P&G could have provided direct access to the system to defendants, but this would not have covered all of the information they sought.[30] P&G would have had to pay over thirty million dollars to obtain all of the information from IRI and even if it had, it would not have had sufficient storage capacity for the data.[31] The court of appeals found that the district court had failed to address the fact that P&G did not "possess" the data and along with the defendants' failure to prove prejudice, reversed the sanctions order.[32]

[11] In another case involving a Lanham Act claim, a plaintiff sought discovery about the defendants' sales of an alleged infringing product.[33] One of those defendants, Wal-Mart, responded with 1,771 pages of Bates-stamped documents that represented a print-out of the tabular view of the raw data within its sales database.[34] Plaintiff claimed that the printouts, with line item data arranged by columns and UPC codes, was "indecipherable" and thereby an insufficient response.[35] The court was

[28] *See Procter & Gamble Co.*, 427 F.3d at 732-33, 735-37.

[29] *See id.* at 739.

[30] *See id.*

[31] *See id*. In 2013, it may seem unbelievable that a major corporation, like P&G would be unable to afford sufficient storage capacity for this data. However, when this case was decided in 1995, the court recognized $30 million as a prohibitive storage cost. *See id.*

[32] *Id.* at 739-41.

[33] *See* Powerhouse Marks, L.L.C. v. Chi Hsin Impex, Inc., No. Civ.A.04CV73923DT, 2006 WL 83477, at *1-2 (E.D. Mich. Jan. 12, 2006).

[34] *See id.* at *1, *3.

[35] *Id.* at *3.

"convinced" that Wal-Mart's burden in deriving the information from the database was "significantly less" than on the plaintiff since Wal-Mart controlled the system.[36] For this reason, the court granted plaintiff's motion to compel a more sufficient response from Wal-Mart.[37]

[12] Finally, in an Americans with Disabilities Act claim, the Equal Employment Opportunity Commission ("EEOC") sought to compel production of portions of the human resources database of a Supervalu and Jewel-Osco, major national food retailers.[38] The EEOC originally sought broad production of information from the human resources database, but narrowed its requests after a meet and confer session to employee hiring, transfer, and termination records, along with job postings for the subject time period.[39] The EEOC premised its request on the defendants' own FRCP 30(b)(6) testimony that "this sort of analysis could be completed" and that defendants' "types of database are designed for this sort of production at minimal expense."[40] Defendants first claimed that they did not have the particular database tool activated in their system to allow them to provide the information requested by the EEOC.[41] Defendants then objected to the scope and burden of the request, claiming that the information would cover over 180 locations and 100,000 employees (when there were only 108 claimants) and that it would take their IT personnel over a week to write the code necessary to obtain the data.[42] The court found that the EEOC had not established that the relevance or

[36] *Id.*

[37] *See id.* at *4.

[38] EEOC v. Supervalu, Inc., No. 09 CV 5637, 2010 WL 5071196, at *1 (N.D. Ill. Dec. 7, 2010).

[39] *Id*. at *6-7.

[40] *Id*. at *6.

[41] *Id.* at *7.

[42] *Id.*

benefit of the information outweighed the burden and expense of producing it and thus denied the motion to compel.[43]

III. CAN A PARTY WAIT TO DEAL WITH STRUCTURED DATA UNTIL THAT INFORMATION HAS BEEN REQUESTED?

[13] The information contained in databases can make the difference between winning and losing a case. The *Sedona Database Principles* makes this statement as a matter of plain fact: "Information contained in databases may be the best source for establishing certain facts in a legal dispute. Information stored in this format also may be useful, if not essential, for analyses such as sorting, calculating, and linking to answer quantitative questions presented in a case."[44]

[14] It is a simple matter to move from the abstract language of the *Sedona Database Principles* to concrete situations. Unstructured data, particularly e-mail, instant messages ("IM"), and typical "office" documents (*i.e.*, Microsoft Word, Excel, and PowerPoint) provides evidence of the communication of activities—who knew what and when. People will write e-mail and text messages to others concerning what they did. Similarly, they will draft documents to memorialize actions that they have taken. In contrast, the structured data in transactional and financial databases provides direct evidence of the action—how, how much, and how often. The financial system will show that money was moved and the time and accounts involved. A transactional application will record the supervisor's approval of the money transfer. Thus, the database systems provide a way to "follow the money" and recreate what happened, even if the communications record is incomplete or, in the case of fraud or shady dealing, deliberately obscured. For this reason, some have called

[43] *Supervalu, Inc.*, 2010 WL 5071196, at *8, *12.

[44] *Sedona Database Principles*, *supra* note 5, at 4.

structured data "forgotten data"—"perhaps the single biggest missed opportunity for defense in e-discovery."[45]

IV. PLANNING FOR DISCOVERY OF STRUCTURED DATA

[15] Databases, especially major, enterprise, or department-level systems, are often highly complex and highly customized. The discovery of structured data typically requires specific expertise with experience in deciphering data structures, relationships, and connections to other systems. The *Sedona Database Principles* is filled with warnings about the need for expert assistance,[46] and it likens the act of trying to handle discovery requests involving structured data without such knowledge as "akin to seeing a thousand-piece jigsaw puzzle without an illustration that shows the final completed puzzle."[47]

[16] Seeking information stored in structured data repositories also requires more planning—and often more efforts at cooperation between the parties—than traditional e-discovery. Parties that do not meet and confer before commencing structured data requests may well find that the court sends them back to square one.[48] Many reasons exist for this

[45] Courtney Fletcher & Liam Ferguson, *E-Discovery: Remembering Forgotten Data*, WALL STREET & TECH. (Oct. 21, 2009), http://www.wallstreetandtech.com/regulatory-compliance/e-discovery-remembering-forgotten-data/220900032.

[46] *See Sedona Database Principles*, *supra* note 5, at 2, 6, 12, 17; *see also* Douglas Herman, *Digital Investigations – Where You Forgot To Look: Why Databases Often Are Overlooked When It Comes Time To Harvest Electronic Data*, METRO. CORP. COUNS., (Aug. 2006), http://www.metrocorpcounsel.com/pdf/2006/August/22.pdf ("To extract data from a relational structure[,] such as a CRM or ERP database, requires specific expertise and a solid understanding of the underlying bases of how these databases work.").

[47] *Sedona Database Principles*, *supra* note 5, at 2.

[48] *See* Rebman v. Follet Higher Educ. Grp., Inc., No. 6:06-CV-1476-ORL-28KRS, 2007 WL 1303031, at *3 (M.D. Fla. May 3, 2007) (Plaintiff's broad request for data from a database with over 200 million records denied by the court as overbroad; court ordered

heightened need for additional proactive planning and discussion, but none may be more pressing than the fact that downstream production requirements will control the early stage EDRM work conducted in Preservation, Collection, and Processing, and even potentially as far back as the critical Identification phase of e-discovery.

[17] It should come as no surprise that the *Sedona Database Principles* places particular emphasis on one of the core principles from the original *Sedona Principles*:

> Sedona Principle 3: The Early "Meet and Confer"
>
> "Parties should confer early in discovery regarding the preservation and production of electronically stored information when these matters are at issue in the litigation and seek to agree on the scope of each party's rights and responsibilities."
>
> > Sedona Principle 3 is especially applicable in the context of database discovery because of the complicated technical and logistical questions raised by the storage of information in databases. Database discovery may entail some of the most expensive and complex discovery in a litigation matter, and meaningful conversations between the parties early in the litigation can substantially reduce confusion and waste of resources.[49]

[18] Challenges to the discovery of information stored in structured data repositories can occur from both opposing parties and litigants. Many

parties to meet and confer under Rule 26(f) to narrow the request and determine the need versus the burden on the defendant).

[49] *Sedona Database Principles*, *supra* note 5, at 8 (quoting *Sedona Principles*, *supra* note 6, at 21).

of the solutions for best using data from databases require the creation of a new view or analysis that differs from the way that the information is used in the ordinary course of business. Responding to structured data requests is likely to require new reports, new extracts directly from the systems, or even entirely new systems to analyze data. Attorneys are often not comfortable with this process, especially since information about how these new views of structured data were created may have to be disclosed to the other side if challenges arise as to the adequacy of the proffered discovery response. Thus, it is critical to complete a full and frank discussion, between *all* stakeholders—each side and each role (Legal, IT, outside expert)—that clearly sets out all expectations before any work begins.

[19] The first issue that practitioners are likely to confront during the e-discovery process involves the specific elements that will be extracted from the database. In some situations, it may be necessary to preserve and collect elements that would not normally be considered "content," such as reports, formulas, pick lists, reports, queries, and the like.[50] For example, FLSA class action litigation often revolves around issues of how companies determined which employees were exempt from overtime and which were non-exempt; formulas within the HR and payroll systems applying these standards become critical.[51] Fraud cases that center around who knew what and when could require the recreation of standard reports and views that were used at the time of the alleged suspicious activity.[52]

[50] *See id.* at 24.

[51] *See, e.g.*, Ojeda-Sanchez v. Bland Farms, LLC, No. CV608-096, 2009 WL 2365976, at *3 (S.D. Ga. July 31, 2009) (requiring production of entire database as "metadata" where the formulas within the system were relevant to the issues in a wage and hour class action); *see also Sedona Database Principles*, *supra* note 5, at 25 illus. iii.

[52] *See, e.g.*, Goshawk Dedicated Ltd. v. Am. Viatical Servs., LLC, No. 1:05CV2343-RWS, 2007 WL 3492762, at *1 (N.D. Ga. Nov. 5, 2007) (requiring production of database in fraud and truth in lending case required despite respondent's claim that it was confidential and "the single greatest asset" of the party because the accuracy of the data and algorithms therein was highly relevant to the claims and defenses of the case).

Such elements will almost certainly require rigorous preservation and collection methods, such as a complete database copy or a restored full back up, as outlined below.

[20] In most cases, practitioners will need to focus solely on database content: the fields and records. With this approach, legal teams must anticipate potential issues as they either use or produce this information. Concerns include: (1) a need for completeness and usability of the data set; (2) availability of the data and technical feasibility of any planned search and retrieval Methods; and (3) cost. Each concern is explored in turn below.

A. A Need for Completeness and Usability of the Data Set

[21] The fact that some of the data within a database may be relevant does not mean that the entire database must be produced. Sedona Database Principle 1: Scope of Discovery clearly speaks to this point: "Absent a specific showing of need or relevance, a requesting party is entitled only to database fields that contain relevant information, not the entire database in which the information resides or the underlying database application or database engine."[53]

[22] Will legal teams require a complete set of data or merely an extensive subset of potentially relevant records? For a small subset of data, a surgical approach will likely suffice. However, if a complete dataset will be required for further analysis, the scope of database preservation, collection, and production will be much more extensive. Date ranges for activity or database information creation may be helpful at this stage.

[23] Does the team require a picture of the information present at a particular point in time? If so, a snapshot of the data or the system will likely accomplish these objectives. To create a historical record, a trend

[53] *Sedona Database Principles*, *supra* note 5, at 21.

line, or to illustrate changes over time, more comprehensive preservation and collection will be required.

B. Availability of the Data and Technical Feasibility of any Planned Search and Retrieval Methods

[24] Structured data systems have a variety of capabilities and technical capacity. Many of the older legacy systems can be very limited in how one can manipulate and export data. Thus, before making any plans—or worse, commit to a regulator or the other side in litigation as to a methodology or deliverable data—it is critical to determine whether the target system includes the necessary capabilities. The answer to this question will vary by the circumstances of each case, but some of the questions highlighted in Comment 2B of the *Sedona Database Principles*[54] provide a good starting point:

- Can a user run searches within the system, other than those built specifically for the intended business uses of the database?[55]

- Will the searches bring back complete information (*i.e.*, all the requested data)?[56]

[54] *See id.* at 27-30.

[55] *See id.* at 28. The problem of database systems designed for particular purposes, which are not accessible in the ways required for discovery, was specifically recognized by the Standing Committee of the Judicial Conference in its September 2005 Report Recommending the Adoption of the 2006 Amendments, as a potential form or not "readily accessible" ESI under Rule 26(b): "[D]atabases that were designed to create certain information in certain ways and that cannot readily create very different kinds or forms of information." REPORT OF JUDICIAL CONFERENCE OF THE UNITED STATES ON RULES OF PRACTICE AND PROCEDURE C-42 (Sept. 2005), *available at* http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/Reports/ST09-2005.pdf [hereinafter *Judicial Conference Report*].

[56] To optimize database performance, some database systems will only index portions of long, free-form text fields—such as the first few hundred characters—so that search results from such systems may not be complete. *See Sedona Database Principles*, *supra* note 5, at 17, 28.

- Is there information stored outside of fielded tables?[57]

- Does the producing party have custody and control of the database, such that it can access the "back end" of the system to export data, create custom reports, or otherwise access the system outside of normal business use?[58]

- Does the system support third party tools that might be more efficient at querying the data?[59]

- Does the system have reporting capabilities?[60]

- Does the system support the creation of custom reports?

[25] The answers to these and other questions will directly impact the extent to which a case team can preserve, collect, and ultimately produce the data stored within a database system. It is crucial that qualified personnel correctly provide this essential foundational information. It may be necessary to support such statements with documented expert evidence. Given a lesser evidentiary showing, the courts have shown little sympathy for such claims, particularly when made by sophisticated corporations.[61]

---

[57] Some database systems use "look up" tables or "drop down" menus to create pre-defined data entry fields which contain information hard-coded into the system itself, not in any searchable fields. *See id*. at 28.

[58] *See id*. at 29. With the increasing popularity of SaaS systems, such as Salesforce.com, the business user of a system may no longer have any access to a system beyond their usual user interface. *Id*.

[59] *See id*. at 6 (IT departments are likely to require extensive and time-consuming testing of any third-party system that would be installed inside the corporation, especially if it would connect to a mission-critical system).

[60] *See id*. at 29.

[61] *See, e.g.*, Zurich Am. Ins. Co. v. Ace Am. Reinsurance Co., No. 05 Civ. 9170 RMB JCF, 2006 WL 3771090 (S.D.N.Y. Dec. 22, 2006); Static Control Components, Inc. v. Lexmark Int'l, Inc., No. 04-84-KSF, 2006 WL 897218 (E.D. Ky. Apr. 5, 2006).

C. Cost

[26] Structured data discovery has the potential to be more costly than "standard" requests. It is imperative that parties have a strong understanding of the potential costs associated with structured data discovery. Courts have become particularly sensitive over recent years to knee-jerk undue burden and cost claims under FRCP 26(b)(2)(B) that lack concrete documented support.[62] This concern is yet one more reason why retaining experienced experts, who can attest to costs encountered in similar situations, may be critical to adequately educate both courts and requesting parties.

V. HANDLING STRUCTURED DATA WITHIN THE EDRM

[27] The Electronic Discovery Reference Model ("EDRM") has come to provide an industry-accepted workflow for e-discovery across the litigation lifecycle. Discovery of structured data can generally proceed within the EDRM framework, though a number of modifications may be required because of the unique requirements inherent in handling this type of ESI. Virtually all structured data projects will require the application of an IT concept known as "ETL," which is the acronym for Extract, Transform, and Load. A good working definition for ETL is:

However, this does not mean that the courts will necessarily unreasonable requests. *See, e.g.*, *In re* Ex Parte Application of Apotex Inc., No. M12-160, 2009 WL 618243 (S.D.N.Y. Mar. 9, 2009) (two weeks before scheduled trial, a party in patent litigation sent a broad subpoena for data to a competitor, involving data from over 30 years ago; court denied the request after the competitor demonstrated the difficulty of obtaining the data).

[62] *See, e.g.*, Cartel Asset Mgmt. v. Ocwen Fin. Corp., No. 01-cv-01644-REB-CBS, 2010 WL 502721 (D. Colo. Feb. 8, 2010) (rejecting claim that ESI was inaccessible due to burdensomeness after respondents failed to provide specific information regarding their storage practices, the number of storage systems that they would need to search, and their capability to retrieve information from those systems).

> ETL is short for extract, transform, load, three database functions that are combined into one tool to pull data out of one database and place it into another database. [Extract] is the process of reading data from a database. [Transform] is the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data. [Load] is the process of writing the data into the target database.[63]

[28] ETL is required in e-discovery for the simple reason that most business-oriented database systems (*e.g.*, Peoplesoft, Cognos, Oracle Financials, specialized procurement software, and SQL databases) are designed to meet specific business needs and do not inherently "speak" to each other. Hence, ETL permits different data formats to be assimilated or aggregated in a unified source for analysis. This saves time querying multiple databases in various coding languages to try to quantify an impact, establishing relationships with the data across systems, and providing meaningful results to counsel and client.

[29] For structured data, a typical workflow involves an ETL overlay of several EDRM phases, beginning with Identification and typically running through Preservation and Collection, and at times into the Processing phase. This process is illustrated in the figure reproduced in the Appendix.

A. Identification

[30] The Identification phase for structured data is likely to require substantially more experience than it normally would for unstructured data systems. Large-scale enterprise database systems, such as Oracle, SAP and PeopleSoft, are highly complicated and customized, requiring advisors

[63] *What is ETL (Extract, Transform, and Reload)?*, WEBOPEDIA, http://www.webopedia.com/TERM/E/ETL.html (last visited Mar. 12, 2013).

with specialized expertise to understand them. This complexity may even be considered a trade secret and thus protected by the software vendor.[64] Even small-scale systems as simple as Microsoft Access databases are often customized and connected to other systems in ways that are both unexpected and poorly documented. Older structured data repositories that fall into the categories of legacy data, obsolete hardware, and retired systems may present particular concerns since the documentation that existed at one time may no longer be available or accurate. Further, the employees who created and maintained these systems may be long gone from the company, having taken with them any institutional knowledge about these systems.

[31] For all of the above reasons, Sedona Principle 6: Responsibilities of Responding Parties is particularly applicable to and significant for the discovery of structured data. Sedona Principle 6 reads: "Responding parties are best situated to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own electronically stored information."[65] The *Sedona Database Principles* further apply this guidance to the discovery of structured data in Database Principle 2: Accessibility and Proportionality, which states: "Due to the differences in the way that information is stored or programmed into a database, not all information in a database may be equally accessible, and a party's request for such information must be analyzed for relevance and proportionality."[66]

[32] However, the fact that a producing party is generally better situated to evaluate methodologies and burdens does not mean that the responding party can and should examine and evaluate such information unilaterally. In accord with the *Sedona Database Principles*' focus on cooperation between the parties, Database Principle 3: Use of Test Queries and Pilot

[64] *See Sedona Principles*, *supra* note 6, at 30.

[65] *Id.* at 38.

[66] *Sedona Database Principles*, *supra* note 5, at 26.

Projects recommends that the parties work together, starting with the sharing of database and system documentation or even going so far as to create test queries and pilot projects. It states: "Requesting and responding parties should use empirical information, such as that generated from test queries and pilot projects, to ascertain the burden to produce information stored in databases and to reach consensus on the scope of discovery."[67]

[33] Key goals in the identification phase should include:

- Determining which systems are likely to include data that might need to be used or produced;

- Establishing the current status and availability of the data, such as whether it is still within live data systems, in legacy systems, in archives, on backup media, offline, legacy or retired systems;[68]

- Locating the data, as many database systems have parts spread out among many physical locations, often in remote server farms or co-location facilities;[69]

- Ascertaining who controls those systems (a vendor, such as SalesForce or other third party, rather than the client/litigant, may actually have possession and day-to-day control over the database itself);

[67] *Id.* at 31.

[68] Legacy and retired systems are commonly found in corporate acquisitions, where an acquired company's IT systems tend to be, at best only partially migrated over to the acquiring company or simply taken offline. There may be no current users or administrators of such systems at the current company. *See id.* at 14; Herman, *supra* note 46 ("Some systems, especially those that are older, may have been grouped together as a result of certain corporate mergers and acquisitions and may not be operating efficiently or may not be stable . . . .").

[69] *See Sedona Database Principles*, *supra* note 5, at 13.

- Understanding the functional purpose of those systems, both for which they were created and potentially for any later purpose or purposes for which they may be currently used;[70]
- Determining the capabilities and limitations of the current system or media holding the data—an important step that will set practical boundaries for how data can be preserved, collected and processed;

- Assessing the costs and burdens of obtaining—and if necessary restoring—the data from its current storage repository; and

- Evaluating the potential benefit of obtaining the data.

[34] Data flow and entity relationship diagrams can be particularly useful in tracking down database connections, assuming the company has taken the time to create such documentation. This documentation augments the more technical documentation involved with *data mapping* and a *data dictionary* or *schema*. Data mapping, which is a list of how enterprise systems interconnect (sometimes prepared as a list, but sometimes created as an actual graphical map),[71] can make the difference between the success and failure of the project. Structured data systems connect to other systems within the enterprise, often to many systems and in surprising ways. Missing those connections can mean missing necessary inputs, outputs, and related or relevant data.

[35] A data dictionary or schema shows the type of data that is in a system, how it is organized and named, and the relationships between that data as it sits in fields and tables.[72] Since structured data systems are often complicated and expensive, these tools tend to have long lives and may have changed purpose or focus over time. As it can be burdensome to

[70] *See id.* at 12.

[71] *See Sedona Glossary*, *supra* note 1, at 13.

[72] *See Data dictionary*, DICTONARY.COM, http://www.dictionary.reference.com/browse/data+dictionary (last visited Mar. 16, 2013).

modify an underlying data table structure, newer data may be stored in repurposed fields or tables that may not be properly named or intended for the current use. Such informal modifications are rarely fully documented unless a conscious (and recent) effort has been made to build a schema. However, as underscored by Comment 1B of the *Sedona Database Principles*, data that could initially appear to be irrelevant may in fact be relevant because of its relationship and connection to other data fields.[73] Thus, it is no surprise that the *Sedona Database Principles* propose that the responding party has a duty to provide the requestor with the information needed to convey a "basic understanding" of the database system.[74]

[36] A final challenge in the identification phase is that the most common users of these structured data systems, the end-users or "customers" who query the substantive information stored in the database, are unlikely to be experienced IT professionals. These users rarely have the time, knowledge, or ability to wade through technically confusing scenarios that a legal case team may pose. A case team must take this into account and plan to interview a mix of end-users and database-knowledgeable IT professionals in order to build a reasonable understanding of a complex structured data repository in active use.

B. Preservation and Collection

[37] One of the most troubling aspects of e-discovery is that ESI has a tendency to disappear unless properly preserved. Backup tapes get recycled, e-mail servers are purged of ex-employee accounts, and hard drives from the laptops of ex-employees are reformatted and reused. Depending on the specific system at issue, some structured data repositories may be even worse in this regard. While much unstructured data is lost due to human action, certain types of common structured data

[73] *Sedona Database Principles*, *supra* note 5, at 23.

[74] *Id.* at 25.

systems are specifically designed to eliminate or overwrite data regularly and automatically, without anyone's direction or oversight.

[38] These repositories stand in contrast to databases comprised of historical information, such as customer relationship management systems, complaint or incident databases, and financial systems used to determine trends, which are typically designed to log all inputted information. In these systems, where one of their intended uses is long-term "data mining" for analytical purposes, the danger that information will disappear is appreciably less.

[39] High volume transactional systems tend to overwrite data or regularly purge old data as the need for historical data is often limited and the volume of data that would build up over time would become prohibitively expensive to store.[75] This problem is well known and the drafters of the 2006 FRCP Amendments who added the rules on ESI specifically noted that "many database programs automatically create, discard, or update information" and "that suspending or interrupting these features can be prohibitively expensive and burdensome."[76] Thus, practitioners assisting in a matter that touches these types of data systems will need to act quickly to preserve this type of system to avoid being left with incomplete data or none at all.[77]

[40] Another unique wrinkle to the discovery of structured data is that the lines between the Preservation and Collection phases tend to blur. For

[75] *See, e.g.*, Procter & Gamble Co. v. Haugen, 427 F.3d 727, 739 (10th Cir. 2005) (finding that the responding party would have to purchase a mainframe computer to download and archive the data at its own facilities or purchase the archival data from the third-party at a great cost).

[76] *Judicial Conference Report*, *supra* note 55, at C-83.

[77] However, even if portions of the data from such overwriting systems have disappeared by the time respondent acts, the court may still require production of what remains. *See, e.g.*, Burkybile v. Mitsubishi Motors Corp., No. 04 C 4932, 2006 WL 3191541, at *4 (N.D. Ill. Oct. 17, 2006).

structured data, the information that is preserved is often exactly what is collected. Most unstructured formats include potentially responsive files that are moved from at-risk locations (laptop hard drives, USB flash drives, unsecured network file stores, e-mail inboxes, *etc*.) to secure, locked down media or formats, pending further analysis. In contrast, non-purging structured data typically needs to be collected from the underlying system to be preserved. Thus, an already deadline-intensive e-discovery process can become more fraught with difficult-to-make and far-ranging early decisions.

[41] It is important, however, to reemphasize that the fact that a database contains relevant information does not mean that the entire system must be locked down under a legal hold. Sedona Conference Principle 5: Duty of Preservation places a practical limit on the expectations of the parties: "The obligation to preserve electronically stored information requires reasonable and good faith efforts to retain information that may be relevant to pending or threatened litigation. However, it is unreasonable to expect parties to take every conceivable step to preserve all potentially relevant electronically stored information."[78] Thus, parties can use a number of different methods to collect and preserve structured data; the choice will be driven not by the impossible expectation of perfection, but by the circumstances of the case and the project scope questions previously discussed in "Planning for Discovery of Structured Data."[79] Each of these collection methodologies has advantages and disadvantages. Improperly applied, some methodologies have the potential to harm the information integrity of the underlying database and therefore, need to be used carefully or may need to be discussed more fully with the requesting party before moving forward.

[78] *Sedona Principles*, *supra* note 6, at 28.

[79] *See supra* Part IV.

1. Forensic Collection of the Live Database

[42] Some disputes may require preservation and production of a complete copy of the database system. For example, this may be necessary where questions exist about the integrity or functionality of the database as a whole or if there is a need to manipulate the data in some way other than just as a historical record.

[43] Collecting an entire database has some advantages, such as in situations where the complete dataset or evidence must be preserved. This method presents the path of least resistance to key issues of data verification and authentication in that data can be verified through MD5 or SHA-1 hash codes to authenticate it as the basis for its admissibility as factual evidence. Complete collection also presents the safest route against spoliation as any changes to a database in active service will not impact the version that was collected and is now out of tinkering hands.

[44] That said, copying an entire structured data repository also has disadvantages when compared to other information collection methodologies. The first disadvantage is cost. Unless the system is small (*e.g.*, desktop computer-based), the sheer size of a data repository may require large amounts of storage media, significant IT investment, and costly disruption to corporate operations. In addition, accessing a collected data repository may require building a comparable hardware and software environment to load, search, and otherwise manipulate the data. Enterprise-level infrastructure for this task is likely to be quite costly, even on a short-term leased basis. For older legacy systems, it might not even be possible to copy the system and even if were possible, duplicating the computer systems on which the information resides might have long become unavailable. Contractual rights may prevent this collection methodology. In the case of databases accessed over "the Cloud," copying the database as a whole is strictly forbidden both by license and deliberately-created technical constraints.[80]

[80] *See, e.g.*, *Conditions of Use*, SORENSON MOLECULAR GENEALOGY FOUNDATION, http://www.smgf.org/terms/jspx (last visited Mar. 11, 2013); *Copyright Information*,

[45] It is important to note that the preservation and collection of an entire database is rarely required for most legal disputes. Most e-discovery requests involve only a subset of structured data. Thus, collecting an entire database to preserve only a small amount of information within it incurs additional time and expense to search, cull, and select data, all of which will have to be done outside of the easy confines of an e-discovery review tool.

2. Restoration of Backups from the Database

[46] Similar in outcome, but potentially less burdensome, disaster recovery backups of a structured data repository may be used to preserve and collect databases. Most organizations have regular business continuity backups of their key systems and it may be less onerous to divert one of these data snapshots than it would be to make a full copy of the live database. However, the same disadvantages apply as making a copy of the system, along with some additional challenges that may make this potential methodology inappropriate in many situations.

[47] Backup media may contain not just data regarding the database at issue, but also data from completely different systems as well. Separating this information will require additional time and expense and may be complicated by data privacy requirements, such as HIPAA, that require the enactment of significant security measures for the removal of data.[81] In addition to these costs, backup media must be restored, again requiring time, IT expertise, and suitable hardware to which the system image can be restored.[82] Finally, backup systems are far from perfect and failure

---

HYPERGEERTZ, http://hypergeertz.jku.at/Geertzcopyrightinformation.htm (last visited March 11, 2013);
Terms of Use, MASSINVESTOR, http://www.massinvestordatabase.com/terms.php (last visited Mar. 11, 2013).

[81] *See infra* Part VI.D.

[82] For these reasons, the *Sedona Database Principles* actively discourage the use of backup tapes as a methodology. *See Sedona Database Principles*, *supra* note 5, at 11.

rates, while not as high as they have been even in the recent past, are still in the words of a highly-respected industry analyst, "not acceptable."[83]

3. Extracting Select Information from the Database

i. All Fields/Data

[48] A more selective and thus more efficient alternative to collecting an entire repository is extracting the substantive data from the system and exporting it into a generic data format that can be read by multiple databases. The success of data collection using this methodology is relatively simple to test, using one of several established techniques. In addition, if the extraction process is handled according to IT industry standard practices and properly documented, authentication should also be relatively straightforward. Capturing a full set of the underlying data permits a case team to defer filtering and culling decisions to a later date, pushing back some expense until it is truly necessary.

[49] Collecting database information through data extraction has some drawbacks. As with other techniques that capture the entire data set, much of what is collected will be irrelevant and will need to be filtered out before any review or production. This can be a lengthy, disruptive, and expensive process. It is important to note that extracting the complete data set does not mean that all of the capabilities of the original database will be available. Much of the value of many database systems stems from the *computed values* and analysis obtained by applying algorithms to source data. Capturing raw data alone is often not adequate to collect this high-value relational information as well.[84] The full extract, transform, and load process may be required to derive potentially critical information.

[83] Dave Russell, *The Broken State of Backup*, GARTNER, 1, 5-6, http://www.cornerstonetelephone.com/sites/default/files/resources/Gartner_-_The_Broken_State_of_Backup_(6-09).pdf (last visited Mar. 23, 2013).

[84] *See Sedona Database Principles*, s*upra* note 5, at 20.

ii. Selected Fields

[50] Because databases typically track much more information than is relevant to a particular legal matter, it may be possible to extract select information stored within it. Such selection can be applied along two axes: (1) limiting data extraction to a subset of database records and selecting them through an appropriate search query; and (2) limiting data extraction to only a subset of fields within a database record. Often, both limitations are applied in the same export. This approach has clear advantages in terms of cost, data volume, and amount of time required to complete the requested data extraction. However, by the same token, leaving behind some of the validating information found in a database field may make the extracted information more difficult to authenticate.

[51] Identifying and extracting the relevant data depends on three things: (1) knowledge of the system; (2) understanding of the matter; and (3) skill at creating queries. Deficiencies in any one of these areas may complicate this effort. In addition, because not all of the data in a database is collected using this methodology, there is some risk if the database has an information purging function built into it. It may not be possible to fix mistakes if the initial selection criteria turn out to be incomplete. Fortunately, when cooperation exists between all participants and parties in the process, this collection methodology can be both efficient and cost-effective for everyone.

iii. Sample Fields (and Potentially Reiterations as Needed)

[52] When the existence or non-existence of potentially relevant information is an open question, a final form of data extraction is to export sample database records. The process can be repeated reiteratively, even incorporating suggestions from the requesting parties. Properly conducted, this approach may permit a structured data repository to be dismissed as a source of potentially relevant information or it may hone the criteria required to identify and extract appropriate information. Either

way, approaching such an investigation cooperatively, rather than unilaterally, may enhance the defensibility of this approach.[85]

[53] Selected sampling incorporates the risk factors that arise when extracting only select information from a database. This approach adds a fourth potential failure point: the need for competence in generating appropriate sample sets and testing them for potential relevance. Because of the highly selective nature of this approach, rigorous documentation is required to answer questions that may arise later as to the adequacy of how this methodology was applied.

4. Reports

i. Using Existing Reports

[54] Existing (*i.e.*, "canned") database reports that are used for business purposes can be a useful first step for collecting structured data. First, the total data volume will be much lower than other methods unless the reports are themselves massive. However, as Comment 1F of the *Sedona Database Principles* highlights, even voluminous reports may still be appropriate to produce even with the inclusion of additional non-responsive information, as this could be the easiest, least expensive, and least burdensome way to obtain and produce the information so long as the producing party is not doing so for any improper purpose.[86] Second, existing reports were created and generated for business purposes and thus have typically been "pre-validated." The accuracy of the information presented has been accepted as accurate and reliable as the basis for business decisions.[87] This can greatly simplify post-production validation and authentication. Third, these reports are typically minimally intrusive for an organization. The report templates and underlying queries have

[85] *See id.* at 31.

[86] *Id.* at 26.

[87] *Id.* at 19.

already been created and used in the ordinary course of business so no custom workflow must be developed. Fourth, especially with respect to Cloud-based/SaaS type proprietary systems, reports may be the only way to retrieve data from a system.

[55] Unfortunately, the use of existing reports is not a perfect collection solution. These reports were designed for specific business needs, not the needs specific to a legal dispute. For this reason, existing reports rarely provide the information that is specifically requested. They typically provide too much or too little. A troubling problem, and one that is less well understood, is the fact that reports tend to not be "pure" output from the system. Many database reports are compilations and aggregations of information that are more than raw information output from stored information. Instead, this raw information may be added, reformatted, or otherwise "tweaked" from the pure source information in the database, sometimes to the point of showing significant deviation from source information. For purposes of validation and authentication, this can create obvious problems.

[56] Often, requesting parties do not automatically accept database reports in lieu of direct discovery of the source database. In addition, it would be unwise to assume that the courts will side with the producing party over this issue without first examining the underlying facts leading to the creation of specific reports.[88]

[88] *See, e.g.*, Margel v. E.G.L. Gem Lab Ltd., No. 04 Civ. 1514(PAC)(HBP), 2008 WL 2224288, at *4-6 (S.D.N.Y. May 29, 2008) (ordering respondent to produce the database as well as the reports from the database because the database was not in the same form, under FRCP 34, as the reports). *But see, e.g.*, EEOC v. Supervalu, Inc., No. 09 CV 5637, 2010 WL 5071196, at *8 (N.D. Ill. Dec. 7, 2010) (rejecting a request that would have required creation of custom report that would have taken two weeks work where requestor could not prove that the relevancy of the data to be obtained was greater than the burden on the respondent).

ii. Creating Customized Reports

[57] Another option for data extraction from structured data repositories is to design a custom report. Custom reports provide greater flexibility than existing reports due to their ability to be limited to relevant data, data fields, and time periods. Custom reports also help to limit inadvertent disclosure of irrelevant data and can even be used on privileged, confidential, or protected personally identifiable information.

[58] As a word of caution, not every system allows for the creation of custom reports, and even when this functionality is available, it may be difficult or expensive to use. Custom reports may face a greater evidentiary hurdle than canned reports used in day-to-day business operations. However, courts have been somewhat more sympathetic to production objections based on the undue burden of creating expensive custom database reports to comply with incoming discovery requests.[89]

5. TIFF Image Snapshots

[59] An older, and now less commonly accepted, way to produce structured data is to capture database output sent to the monitor or to reports and to render these "snapshots" to TIFF image. This creates an easily preserved form that can be Bates-stamped and for which authenticity can easily be stipulated.[90] While appropriate in some situations, this production method has fallen out of favor compared to

[89] *See, e.g.*, Soto v. Genentech, Inc., No. 08-60331-CIV, 2008 WL 4621832, at *12 (S.D. Fla. Oct. 17, 2008) (allowing responding party to produce detailed log of data contents in lieu of creation of custom reports that would have required approximately 64 hours of work); *see also* Getty Props. Corp. v. Raceway Petroleum, Inc., No. Civ. A. 99-CV-4395DMC, 2005 WL 1412134, at *4 (D.N.J. June 14, 2005).

[90] This method was originally suggested by Thomas Allman in an early and seminal review of the then brand-new 2006 ESI FRCP Amendments. *See* Thomas Y. Allman, *Managing Preservation Obligations After The 2006 Federal E-Discovery Amendments*, 13 RICH. J.L. & TECH.. 9, 48 (2007), *available at* http://law.richmond.edu/jolt/v13i3/article9.pdf.

other alternatives since it tends to reduce the fielded nature of the underlying data, thereby turning structured data into flat, inflexible unstructured documents that may or may not contain searchable text. That being said, certain database systems have such limited data output capabilities that capture of data in this manner may be one of the only options currently available.

6. Direct Access to the System

[60] A final method for producing information from a database is to simply let the requesting party or its expert have direct access to that system to run its own queries or reports. However, most litigants highly disfavor this method as it allows the opposing party potential access to privileged and confidential information within the database. Courts that have addressed this situation have tended to be receptive to such concerns, requiring that limits be set.[91] This direct access approach also has significant potential to disrupt in-house IT infrastructure and staff who are likely to be unhappy at opening a controlled organization's asset to interlopers. Indeed, the Committee Notes to the 2006 Amendments to FRCP 34 make it quite clear:

> The addition of testing and sampling to Rule 34(a) with regard to documents and electronically stored information is not meant to create a routine right of direct access to a party's electronic information system, although such access might be justified in some circumstances. Courts should guard against undue intrusiveness resulting from inspecting or testing such systems.[92]

[61] In addition, granting outsiders access to data repositories containing certain personally identifiable information may violate data

[91] *See, e.g.*, *In re* Ford Motor Co., 345 F.3d 1315, 1316-17 (11th Cir. 2003); *In re* Lowe's Cos., Inc., 134 S.W.3d 876, 879-80 (Tex. App. 2004).

[92] FED. R. CIV. P. 34(a) advisory committee's note.

privacy laws and create significant (albeit unrelated) liability for the producing party. For these reasons, direct access to databases and other such systems tends to be granted over objection "only in extraordinary circumstances."[93]

[62] No matter what process is used to preserve and collect a database, proper documentation and testing is critical as many of these processes are complicated and mistakes can occur. Proper documentation and a record of testing will help to demonstrate good faith efforts if these procedures are later called into question.[94]

C. ECA and Processing

[63] Once the data has been extracted from its repository, it typically undergoes further transformation so that it can be used in the investigation or litigation context prior to attorneys' review for substance. For loose documents, litigants typically apply early case assessment techniques, such as key word or concept filtering, to reduce the data volume.[95] Unfortunately, such techniques do not apply well to structured data, as this information is largely centered around transactions rather than words and phrases. Properly processing and limiting the volume of such systems can instead profile the transactions using specific fields, dates, and general ledger codes. A strong understanding of the system at hand becomes even more important in such situations.

[64] Traditional culling methods may be more helpful when the extracted data includes free-form text entry fields such as "comment" fields. Even here, though, because the unified extracted data exists as a single mass of (fielded) information, culling this glob of information can

[93] *Sedona Database Principles*, *supra* note 5, at 16.

[94] *Id*. at 17.

[95] *See id.* at 3.

raise evidentiary challenges unless all changes are well documented and ideally, negotiated at least in principle with the requesting party.

D. Review and Analysis

[65] Once the data has been processed and preliminary analytics have been applied, it may still need to be reviewed for responsiveness and privilege. Some structured data can be managed within standard review platforms, especially flat-file reports and data tables rendered as Microsoft Excel spreadsheets. On the other hand, data extracted from enterprise-grade relational databases cannot be loaded into a review platform with any genuine hope of validly reviewing this information. As described in the *Sedona Database Principles*:

> Analyzing email messages and discrete electronic files typically involves a team (sometimes a large team) of reviewers and takes place through a document review platform. Such review and analytical tools, however, are a poor fit for the matrices of information found in tables of extracted database information. Instead, review of this information may require technically sophisticated analysts to query the data and extract the meaning of its aggregated information.[96]

[66] A more straightforward approach to reviewing structured data looks not to the data's abstract relevance, but instead to the significance of its data values. Certain field information, such as protected private information, may be redacted or stripped, but this is the closest analogy to the parallel review process that takes place in a document review platform. Otherwise, extracted data is manipulated, queried, and explored. In addition, once protected and privileged data fields are removed from extracted structured data, no further attorney review of individual data fields is typically required.

[96] *Id.* at 10.

[67] When the content of individual data fields, such as notes or memo fields, require attorney review, the review paradigm must be further adjusted. Such a review is complicated by the fact that the information that requires review tends to be stored in a structured manner, but contains unstructured data, such as free text that lacks parameter constraints on length or format. Technical specialists are typically enlisted to develop secure web-based database review tools that present this information in a reviewable format for redaction purposes. Certain profiling and culling methods can be employed to reduce the overall volume of information that requires attorney review, but generally, some "eyes-on" attorney review will be required.

E. Production

[68] Extraction and Transformation processes largely set the production of structured data. Information that has been shed as a by-product of transformation processes may now be non-replicable since many forms of extraction do not allow conversion back "upstream." You cannot, for example, extract data as reports and then reconstitute the data to produce it as a complete database. Such is the reason that Sedona Database Principle 6: Form of Production reminds us that: "The way in which a requesting party intends to use database information is an important factor in determining an appropriate format of production.[97] Comment 6A of the *Sedona Database Principles* takes this even further by underscoring that "it may be impossible for a responding party to take appropriate steps to provide database information in a reasonably useful format if it has no idea how the requesting party intends to use it."[98]

[69] Even if the parties do not avail themselves of the warnings of the *Sedona Principles* and the *Sedona Database Principles* and decline to work together to determine a reasonably usable production format, this

[97] *Id.* at 36.

[98] *Sedona Database Principles*, s*upra* note 5, at 36.

lack of agreement does not mean that parties are free to produce data in any format they choose. FRCP 34(b)(2)(E) requires:

> (E) Producing the Documents or Electronically Stored Information. Unless otherwise stipulated or ordered by the court, these procedures apply to producing documents or electronically stored information:
>
> > (i) A party must produce documents as they are kept in the usual course of business or must organize and label them to correspond to the categories in the request;
> > (ii) If a request does not specify a form for producing electronically stored information, a party must produce it in a form or forms in which it is ordinarily maintained or in a reasonably usable form or forms; and
> > (iii) A party need not produce the same electronically stored information in more than one form.[99]

[70] Courts have shown that they will be alert to production formats that are not usable.[100] Courts can also order parties to produce data in particular formats even if it requires the creation of entirely new data sets.[101] However, at the same time, the full cost of producing structured data does not always fall entirely on the producing party. In some circumstances, a requesting party may be required to bear the burden and

[99] FED. R. CIV. P. 34(b)(2)(E).

[100] *See, e.g.*, Powerhouse Marks, L.L.C. v. Chi Hsin Impex, Inc., No. Civ.A.04CV73923DT, 2006 WL 83477 (E.D. Mich. Jan. 12, 2006) (showing that the defendant produced financial database by delivering 1,771 Bates stamped pages of print outs of the raw field data).

[101] *See, e.g.*, *In re* eBay Seller Antitrust Litig., No. C 07–1882 JF, 2009 WL 2524502, at *2 (N.D. Cal. Aug. 17, 2009) (ordering eBay to create a new data set to produce additional responsive documents, despite its Senior Director of Data Warehouse Development's representation that "it would take an engineer forty-eight hours to format a query, at a cost of $7,200" in order to do so).

expense of some degree of transformation of the data from the producing party so long as the format of the production was in fact reasonable.[102]

[71] The *Sedona Principles* echo the concerns of the courts in Principle 12: Form of Production and Metadata:

> Absent party agreement or court order specifying the form or forms of production, production should be made in the form or forms in which the information is ordinarily maintained or in a reasonably usable form, taking into account the need to produce reasonably accessible metadata that will enable the receiving party to have the same ability to access, search, and display the information as the producing party where appropriate or necessary in light of the nature of the information and needs of the case.[103]

[72] Difficulties can arise when an opposing party requests that structured data be produced in "native format"—that is, the original file format in which producing party keeps the ESI. Courts have sometimes shown an un-nuanced willingness to enforce general demands for native format production if it is properly and timely requested, or even if that is lacking, if good cause can be shown[104] or absent a showing of undue burden or hardship.[105] At times, the courts have even required such native

[102] *See Sedona Database Principles*, s*upra* note 5, at 37.

[103] *Sedona Principles*, s*upra* note 6, at 60.

[104] *See, e.g.*, *In re* Netbank Sec. Litig., 259 F.R.D. 656, 681-82, 683 (N.D. Ga. 2009); Hagenbuch v. 3B6 Sistemi Elettronici Industriali S.R.L., No. 04 C 3109, 2006 WL 665005, at *3-4 (N.D. Ill. Mar. 8, 2006).

[105] *See, e.g.*, Camesi v. Univ. Pittsburgh Med. Ctr., No. 09–85J, 2010 WL 2104639, at *7 (W.D. Pa. May 24, 2010); *see also, e.g.*, Chevron Corp. v. Stratus Consulting, Inc., No. 10–cv–00047-MSK-MEH, 2010 WL 3489922, at *2-4 (D. Colo. Aug. 31, 2010).

file productions from database systems.[106] Many parties indirectly request this by requesting production of "the entire database."[107]

[73] Unfortunately, a "native file" production for structured data can present a number of difficult and unique problems. First, and most obvious, the proprietary database format in which relevant data is stored may not be readable and thus, not "reasonably usable" to the requesting party. Handing over to the other side a complete copy of a database system, particularly a world-class enterprise system, is also not a sufficient solution. The recipient may well need to obtain a licensed copy of the system—a potentially very expensive proposition in the case of high-end database systems—or a near impossible proposition in the case of legacy or obsolete systems that are no longer commercially available (even as they remain protected by copyright and license restrictions from free copying). Even if a license for the system can be obtained, installation of the system could take weeks or months and success is not always a given.[108] Finally, even once such hurdles are successfully overcome, the very first use or view of a copied database system is likely to change the information therein, as such systems typically have tracking capabilities that are difficult or even impossible to turn off, making the copy no longer an accurate copy.[109]

[74] For all of these reasons, more transformative production formats, which change the data from the way it is stored in the ordinary course of

[106] *See, e.g.*, Ojeda-Sanchez v. Bland Farms, LLC, No. CV608-096, 2009 WL 2365976, at *3 (S.D. Ga. July 31, 2009); Perez-Farias v. Global Horizons, Inc., No. CV-05-3061-MWL, 2007 WL 991747, at *3 (E.D. Wash. Mar. 30, 2007).

[107] Michael Spencer & Diana Fasching, *Less Production Can be More in Database Discovery*, L. TECH. NEWS, Oct. 26, 2012.

[108] Even highly sophisticated corporations have at times experienced disastrous failures in attempting to install and use high-end database systems. *See* Ericka Chickowski, *Five ERP Disasters Explained*, BASELINE MAG., Apr. 4, 2009, *available at* http://www.baselinemag.com/c/a/ERP/Five-ERP-Disasters-Explained-878312/.

[109] *See Sedona Principles*, *supra* note 6, at 5.

business, have become a commonly accepted discovery practice.[110] In addition, a strong argument can be made that the fielded nature of the raw data, not the proprietary container in which it is stored, is the essential element that provides "native format" flexibility to this information. If this argument is accepted, further transformation of the data may provide increased accessibility without compromising essential functionality.

VI. ISSUES BEYOND THE EDRM

[75] Because structured data does not fit squarely within an EDRM that was implicitly designed for unstructured data types, it should come as no surprise that additional issues often arise in working with structured data in discovery.

A. Custody and Control

[76] A respondent in discovery is only required to turn over what is in their possession, custody, and control.[111] This obligation extends to traditional materials and ESI alike as well as to unstructured and structured data alike. Complex databases, however, can challenge the issue of where data is stored and the extent to which it is "owned" by the content creator. For example, a database may be housed entirely within a corporation and serviced by company IT professionals, so there would be no possession, custody, or control issue. However, when the database is provided by a service provider, questions about information ownership can and do arise. The licensing provisions for many Cloud-based SaaS providers hold that while information entered into the outsourced database may be the exclusive property of the database service client, many of the internal database elements that create relationships between this client

[110] *See id.* at 7.

[111] *See* Tomlinson v. El Paso Corp., 245 F.R.D. 474, 477 (D. Colo. 2007) (requiring a party to turn over data from third-party database of ERISA information because ERISA created clear duties for the employer that negated any claim that such third party data could not be within its possession, custody or control).

provided data are proprietary to the point that a client does not have permission to view these relationships, much less export them in response to a discovery request.[112] As a consequence, the "owner" of information in these systems—the SaaS client—may not have custody or control over a portion of the ESI that it would have to provide if it hosted the database itself.

B. Verifying that the Data Collected is Accurate

[77] Structured data has the unusual property of appearing accurate and precise, even if the substantive information that the database reports is riddled with errors. This issue can occur because the precision of a database search query or report is separate and distinct from the way in which the source data was created or entered into the system. For example, operators at a call center may be asked to enter their recollections and remarks about customer questions and complaints. This information is likely entered quickly as the operators focus on handling as many calls as possible during their shift and it may contain errors. Yet, when this same information appears in a database report, it is likely to have the appearance of an accurate and truthful statement.

[78] Sedona Database Principle 5: Data Integrity, Authenticity, and Admissibility considers this issue: "Verifying information that has been correctly exported from a larger database or repository is a separate analysis from establishing the accuracy, authenticity, or admissibility of the substantive information contained within the data."[113] Thus, in working with structured data, many practitioners have found it useful to separate these two competing questions about "accuracy." It is possible to validate the accuracy of a mechanical data export. For example, certain reference fields or reference values can be exported with the substantive data and those values verified against the source information in the

[112] *See* Alberto G. Araiza, *Electronic Discovery in the Cloud*, 2011 DUKE L. & TECH. REV. 8, 33 (2011).

[113] *Sedona Database Principles*, *supra* note 5, at 34.

database itself. Even something as simple as comparing the number of database records exported against the number of database records returned by a search query is a step in this direction.

[79] Conversely, practitioners can reserve the right to further challenge the accuracy of the information contained within a structured data repository. In evidentiary terms, the authenticity of the information—that is to say, the information was accurately exported from a database—can be the subject of a stipulation, but the truthful nature of the information remains subject to standard challenges as to hearsay and general reliability.[114]

C. Validating Structured Data so that It Can Be Admissible as Substantive Evidence

[80] Validating structured data is an important consideration when working with this form of ESI. As noted previously, many practitioners are able to find common ground and negotiate a stipulation that ESI has been accurately exported or copied from the source database. Authenticity can be mechanically established even though the exported form of the data is unlikely to be identical to the way that the structured data was maintained inside a larger database. The *Sedona Database Principles* recognize and address this problem, in Principle 4: Validation: "A responding party must use reasonable measures to validate ESI collected from database systems to ensure completeness and accuracy of the data acquisition."[115]

[81] The larger issue, though, is finding a consistent workflow for establishing the reliability of structured data so that it may be admissible for the truth of the information contained therein. Because structured data is typically exchanged in the form of data exports or reports, at least one court has found that the business record exception to the hearsay rule is

[114] *See* Lorraine v. Markel Am. Ins. Co., 241 F.R.D. 534, 538 (D. Md. 2007).

[115] *Sedona Database Principles*, *supra* note 5, at 32.

inapplicable as grounds for admitting this information for the truth of the matter asserted.[116] In the case of *Vinhee*, the court required a detailed showing of how information was entered into a database, including all efforts to identify and correct errors.[117] The court further required additional foundation about how the underlying database was managed.[118]

[82] A majority of other courts have imposed a less onerous set of requirements to admit extracted structured data for the truth of the matter concerned.[119] A key point of argument remains the degree to which substantive information entered into a database has been validated as accurate near or at the time of its creation as structured data. Systems that include such validation will have their information more easily ruled admissible than more open and less regulated databases. In such cases, courts may begin to look at some of the *Vinhee* factors as additional extrinsic evidence required to lay a sufficient evidentiary foundation.

D. Privacy

[83] There are many types of database systems that contain vast amounts of private and personally-identifiable information ("PII") such as HR systems, financial systems, healthcare systems, and customer transaction systems to name a few. PII resides in some unexpected databases that most would not expect to contain confidential PII. Web-logging systems, for example, capture unique IP addresses that could be used to track down the identity and location of users. Such protected

[116] *See, e.g.*, *In re* Vee Vinhee, 336 B.R. 437, 447-49 (B.A.P. 9th Cir. 2005).

[117] *Id.* at 448-49.

[118] *Id.* at 448.

[119] *Compare* R.I. Managed Eye Care, Inc. v. Blue Cross & Blue Shield of R.I., 996 A.2d 684, 691 (R.I. 2010) (reiterating a four part test for determining the admissibility of business records under the hearsay rule), *with In re Vee Vinhee*, 336 B.R. at 446 (defining an eleven part test for determining the admissibility of electronic records under the hearsay rule).

information will need to be identified and redacted prior to release of this data to a requesting party. On the plus side, the same analytical measures that can assist with the extraction of the data can often also be used to locate and redact the confidential data, whether by removing it or replacing it with dummy data. However, while such systems cannot always be perfect, many privacy laws are written with such perfection in mind so as to be rather unforgiving even as towards minor violations. Thus, the parties are advised to carefully discuss putting into place protocols, potentially including protective orders, against the possibility of the inadvertent disclosure of PII.[120]

[84] Unfortunately, that is not the end to the potential problems. Because database systems tend to be distributed, portions of a system or systems to which it connects may well physically be located across jurisdictions, such as the European Union, that have strict privacy regulations.[121] Other jurisdictions may not be concerned with the physical location of the data, but instead as to whether the data subjects—those whose information has been collected and stored—live within that jurisdiction.[122] The penalties for violations of these laws and regulations can be severe, so careful legal consideration of the issues before taking action is well advised.[123]

---

[120] *See Sedona Database Principles*, *supra* note 5, at 8-9.

[121] *See generally* Council Directive 90/46/EC, 1995 O.J. (L 281) 39-45 (defining specific privacy protections to be afforded to personal information).

[122] *See, e.g.*, U.S. Dept. of Commerce, *Safe Harbor Privacy Principles*, EXPORT.GOV (July 21, 2000), http://export.gov/safeharbor/eu/eg_main_018475.asp (defining protections for U.S. citizens' data exported to the European Union); *see also* Commission Decision 2000/520/EC, 2000 O.J. (L 215) 7-9 (accepting U.S. Safe Harbor Privacy Principles).

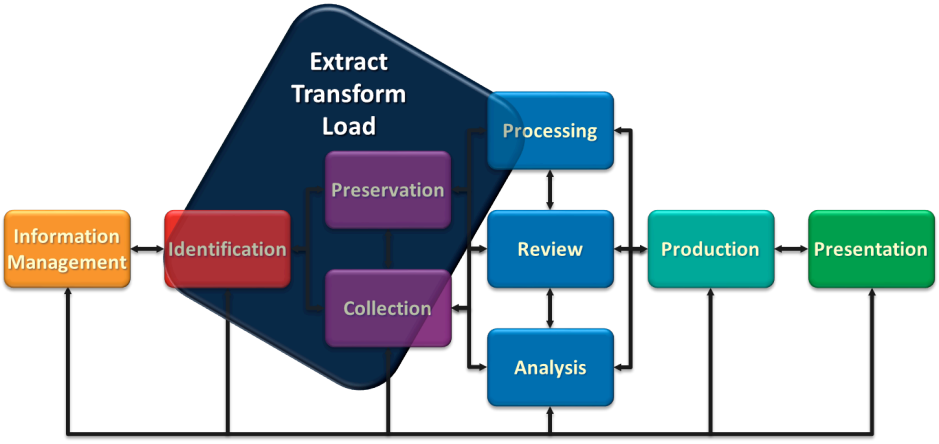
[123] *See, e.g.*, DATENSCHUTZGESETZ 2000 [DSG 2000] BUNDESGESETZBLATT [BGBL] No. 165/1999, §§ 51-52 (Austria) (imposing up to a year in prison and 25,000 Euro fine per violation).

VII. CONCLUSION

[85] Dealing with structured data in e-discovery is something that should neither be ignored nor treated lightly. A case team may be required to handle structured data because an investigator, regulator or the opposing party requests it, or a case team may need to deal with it just to try to understand and prove its case. Situations will arise where the proper expert use of structured data is the best or the only way "to follow the money" and figure out what actually happened. When that situation arises, case teams are likely to need expert assistance to handle the myriad of issues both technical and legal, within the EDRM, and without.

APPENDIX

ETL As Applied to the EDRM Model



*Derived from the Electronic Discovery Reference Model v 2.0, which are used under See Creative Commons Attribution 3.0 United States License. | © 2005-2012 EDRM, LLC.

The Sedona Conference® Database Principles Addressing the Preservation and Production of Databases and Database Information in Civil Litigation[124]

Principle 1: Scope of Discovery

Absent a specific showing of need or relevance, a requesting party is entitled only to database fields that contain relevant information, not the entire database in which the information resides or the underlying database application or database engine.

Principle 2: Accessibility and Proportionality

Due to the differences in the way that information is stored or programmed into a database, not all information in a database may be equally accessible, and a party's request for such information must be analyzed for relevance and proportionality.

Principle 3: Use of Test Queries and Pilot Projects

Requesting and responding parties should use empirical information, such as that generated from test queries and pilot projects, to ascertain the burden to produce information stored in databases and to reach consensus on the scope of discovery.

Principle 4: Validation

A responding party must use reasonable measures to validate ESI collected from database systems to ensure completeness and accuracy of the data acquisition.

Principle 5: Data Integrity, Authenticity, and Admissibility

Verifying information that has been correctly exported from a larger database or repository is a separate analysis from establishing the accuracy, authenticity, or admissibility of the substantive information contained within the data.

[124] *Sedona Database Principles*, *supra* note 5, 21-38.

Principle 6: Form of Production

The way in which a requesting party intends to use database information is an important factor in determining an appropriate format of production.