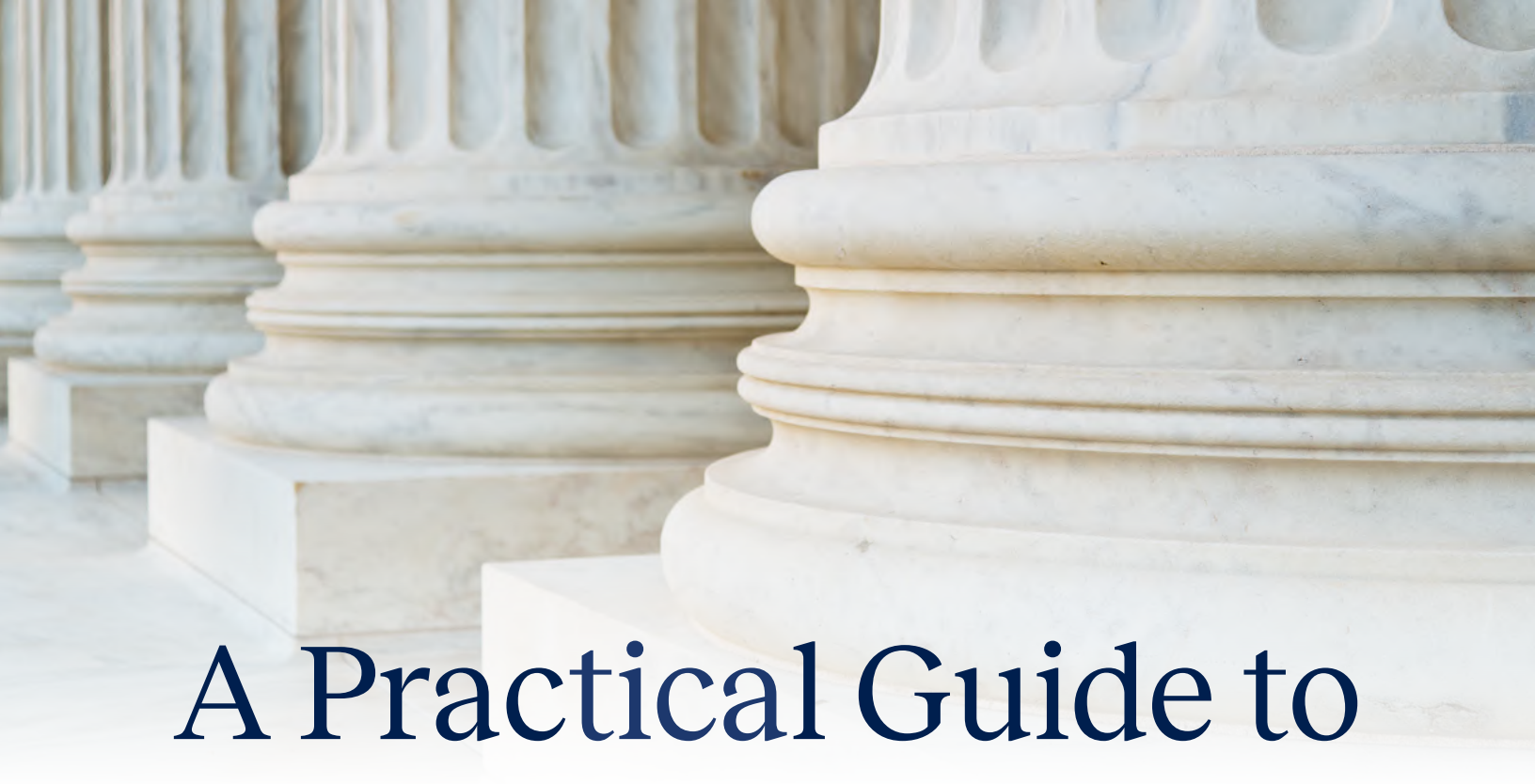
A Practical Guide to eDiscovery

Consilio
Advanced
Learning Institute



A Practical Guide to eDiscovery

by Matthew Verga, et al.

Consilio
Advanced
Learning Institute

Book Cover by Annie Malloy and Harry Woo

Illustrations by Harry Woo

1st edition 2025

Disclaimers

The information provided in this publication does not, and is not intended to, constitute legal advice; instead, all information, content, and materials available in this publication are provided for general informational purposes only. While efforts to provide the most recently available information were made, information in this publication may not constitute the most up-to-date legal or other information.This publication contains links to third-party websites. Such links are only for the convenience of the reader; Consilio does not recommend or endorse the contents of the third-party sites.

Readers of this publication should contact their attorney to obtain advice with respect to any particular legal matter. No reader of this publication should act or refrain from acting on the basis of information in this book without first seeking legal advice from counsel in the relevant jurisdiction. Only your individual attorney can provide assurances that the information contained herein – and your interpretation of it – is applicable or appropriate to your particular situation.

Use of this publication, or any of the links or resources contained within, does not create an attorney-client relationship between the reader and the author or Consilio. All liability with respect to actions taken or not taken based on the contents of this publication is expressly disclaimed. The content of this publication is provided "as is." No representations are made that the content is error-free.

Consilio
Advanced
Learning Institute

About this Book

This book focuses on domestic practice within the United States, primarily civil practice in the federal courts. This book is intended to function in two ways: (1) as a text book that can be worked through linearly to learn about eDiscovery and (2) as a desk reference to which practitioners can turn for guidance about their eDiscovery matters.

This book is organized into three units. Unit 1 – eDiscovery Fundamentals covers the fundamentals practitioners need to know about each phase of an eDiscovery project, from identification to production. Unit 2 – Intermediate eDiscovery provides deeper dives into key eDiscovery skills and challenges for those looking to move beyond the fundamentals. Unit 3 – Advanced eDiscovery explores more challenging subjects, including newer ESI sources and more nuanced legal and procedural issues.

In its electronic form, this book includes hyperlinks directly to sources, so that readers can immediately access and review the sources themselves. Full citations are also provided in the footnotes.

Foreword

In my role as the head of Global Strategic Client Experience for Consilio, I've learned that one perennial client need is the need for continuing legal education – not just the credits required by the bar, but the up-to-date knowledge they can use to improve their projects, their careers, and their organizations. Technology and the law are both continually evolving, making it a constant challenge for practitioners of all types to stay abreast of the latest best practices.

To help practitioners meet this challenge, Consilio launched the Consilio Advanced Learning Institute in 2021, drawing on our deep bench of subject-matter experts and their hundreds of years of combined experience to provide you with the knowledge you need when you need it. Over the past four years, the Institute has provided thousands of accredited CLE hours, along with countless practice guides and white papers, to attorneys from all jurisdictions and across all practice areas.

Now, I am pleased to have the opportunity to introduce you to the latest and largest addition yet to our ever-growing library of educational articles, practice guides, and whitepapers: **A Practical Guide to eDiscovery**. This book reflects the expertise and efforts of many Consilio team members, and we hope you find it to be a valuable tool in your eDiscovery toolkit for years to come.

-**Michael Pontrelli**, Client Experience Officer

Lead Author

Matthew Verga

Matthew is an attorney, consultant, eDiscovery expert, and educator proficient at leveraging his legal experience, his technical knowledge, and his communication skills to make complex legal practice topics accessible to diverse audiences. A twenty-year legal industry veteran, Matthew has worked across every phase of the EDRM and at every level, from the project trenches to enterprise program design. As Director of Education for Consilio, he oversees the activities of the Consilio Advanced Learning Institute, where he works to produce engaging educational resources and programs to empower practitioners with knowledge they can use to improve their projects, their careers, and their organizations.

Co-Authors

Xavier Diokno

Xavier has a bachelor's degree in computer science, a master's degree in computer science, and a juris doctor degree, and prior to becoming an attorney, Xavier worked in the information technology industry for ten years in database administration and software development. For more than a decade, Xavier was part of Consilio's Data Analytics group, where he oversaw the team's tripling in size, as well as numerous large-scale projects involving TAR and novel analytics research. He now applies his technical and legal experience to oversee Consilio's Innovation initiatives, including researching new technologies like AI and developing their application to legal services.

Jon Fowler

Jon is VP and CISO for Consilio. He is a subject matter expert on eDiscovery and digital forensics and advises clients on all aspects of the EDRM, from effective information governance strategies to document production and presentation. He previously headed up Consilio's global digital forensics team, overseeing all data preservation, collection, and forensic examination operations. He has experience as an expert witness in computer forensics in federal and state courts, and has prepared multiple expert reports, affidavits, and statements of fact for clients. Additionally, Jon serves as Adjunct Professor in the graduate program in Computer Forensics at George Mason University.

Mark Garnett

Mark is responsible for managing forensic teams across the US, UK, EU, and APAC regions, including the delivery of all forensic collection, forensic data analysis, discovery consulting, and expert witness services to law firm and corporate clients. Mark specializes in managing teams responsible for collecting and analysing digital evidence, data recovery, electronic evidence preservation, eDiscovery, and expert reporting. Mark is a qualified investigator, eDiscovery and forensic technology practitioner with 14 years' experience as a Detective in the Queensland Police Service and 20 years of eDiscovery and forensic experience, six of which were with a "Big Four" forensic practice in Australia.

Maureen O'Neill

Drawing on more than 25 years of experience providing dedicated client service as an attorney and consultant, Maureen engages with Consilio's customers to ensure a best-in-class client experience. She works with a cross-functional Client Experience Leadership Council to develop best practices that align with the company's commitment to service excellence, and she helps clients find practical, strategic solutions to a range of challenges, including those associated with conducting cross-border discovery. Maureen also serves as Consilio's Diversity & Inclusion Officer. She leads the development and implementation of the company's strategies, policies, and programs for ensuring a diverse, inclusive, and equitable workplace.

Contributors

Sophie Beattie
Vice President, DFES

Mike Gutierrez
Senior Director, DFES

Lorraine Moise
Senior Manager, DFES

Contents

Unit 1

eDiscovery Fundamentals

Chapter 1 - The Evolving Duty of Technology Competence

In discovery specifically, and in legal practice generally, the role of electronically-stored information (ESI) and new technology has grown exponentially over the past decade, as new sources have proliferated, new tools have become normalized, and new communication channels have supplanted the old.

Chapter 2 - In the Beginning: Identification and Preservation Fundamentals

Identification and preservation are the first and most fundamental phases of an electronic discovery effort. The duty of (identification and) preservation is a foundational concept in our legal system that grows out of the common law concept of "spoliation," which is nearly 300 years old.

Chapter 3 - The Grand Scavenger Hunt: Collection Fundamentals

Since ESI has become the norm in discovery, competence with technology has become an essential part of being an effective legal practitioner, and understanding the technology fundamentals of collection has become essential to fulfilling a lawyer's duty of technology competence.

Chapter 4 - Time to Make the Donuts: Processing Fundamentals

The range of potential ESI sources is continually multiplying and diversifying. Processing is how we work with that diverse range of materials without using as many different pieces of software as there are types of sources and how we enable searching and document identification across different source types.

Chapter 5 - Clearing the Fog of War: ECA Fundamentals

The fog of war is apt shorthand for the state of uncertainty that exists early in a new legal matter: What are the facts? What are the risks? What evidence exists, and what does it show? Early case assessment (ECA) is how we start to answer those questions.

Chapter 6 - The Main Event: Review Fundamentals

Document review is typically the most expensive phase of a discovery project, even with the sophisticated tools and techniques available today. Past studies have attributed more than half of discovery costs to review.

Chapter 7 - The Final Countdown: Production Fundamentals

Production is another discovery activity, like collection and processing, in which technical decisions can have logistical and legal effects. For this reason, it is important for practitioners to understand the fundamentals of production.

Chapter 1

The Evolving Duty of Technology Competence for eDiscovery

About this Chapter

In this chapter, we will discuss various aspects of lawyers' duty of technology competence for eDiscovery and how to fulfill them, using the California approach as a model.

1.1 THE EVOLVING DUTY OF TECHNOLOGY COMPETENCE

In discovery specifically, and in legal practice generally, the role of electronically-stored information (ESI) and new technology has grown exponentially over the past decade, as new sources have proliferated, as new tools have become normalized, and as new communication channels have supplanted the old. As a result, it has become a practical reality that effective legal practice and effective discovery requires some level of technology literacy and competence. Since 2012, that practical reality has transformed into a formal requirement, which may be "a very scary wake-up call for some lawyers."[1]

1.1.1 A Formal Duty of Technology Competence

In August 2012, the American Bar Association (ABA) implemented changes[2] to its Model Rules of Professional Conduct, which most state abars look to as a model for their own. Among the changes implemented was a change to make the need for technology competence explicit.

Model Rule of Professional Conduct 1.1[3] establishes a lawyer's general duty of competence in their work, which is the foundational requirement of professional practice:

> A lawyer shall provide competent representation to a client. Competent representation requires the legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation.

The last Comment to that rule[4] covers "Maintaining Competence" over time through continuing legal education (CLE), individual study, and other efforts. The change revised that comment to add technology as an explicit focus[5]:

> To maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, **including the benefits and risks associated with relevant technology**, engage in continuing study and education and comply with all continuing legal education requirements to which the lawyer is subject. [emphasis added]

Although this change was spurred in large part by the rapid rise of ESI and eDiscovery, it is not limited to just that area[6]:

> Broadly speaking, there are five realms of technology competence reasonably necessary for many engagements:
>
> - safeguarding client information
> - ediscovery, including the preservation, review and production of ESI . . .

[1]Victoria Hudgins, States Require Lawyers to Have Tech Competency, But Observers See Some Struggling, LEGALTECH NEWS, https://www.law.com/legaltechnews/2018/10/25/states-require-lawyers-to-have-tech-competency-but-observers-see-some-struggling/ (Oct. 25, 2018), available at https://www.yahoo.com/news/states-require-lawyers-tech-competency-160028982.html.
[2]Debra Cassens Weiss, Lawyers Have Duty to Stay Current on Technology's Risks and Benefits, New Model Ethics Comment Says, ABA JOURNAL, http://www.abajournal.com/news/article/lawyers_have_duty_to_stay_current_on_technologys_risks_and_benefits/ (Aug. 6, 2012).
[3]ABA Model Rules of Prof'l Conduct R. 1.1 (2021), available at https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_1_competence.html.
[4]ABA Model Rules of Prof'l Conduct R. 1.1, Cmt. 8 (2021), available at https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_1_competence/comment_on_rule_1_1.html.
[5]Aug. 2012 Amends. to ABA Model Rules of Prof'l Conduct (2012), available at http://www.abajournal.com/files/20120808_house_action_compilation_redline_105a-f.pdf.
[6]Steven M. Puiszis, Perspective: Technology Brings a New Definition of Competency, BLOOMBERG LAW, https://news.bloomberglaw.com/business-and-practice/perspective-technology-brings-a-new-definition-of-competency (Apr. 12, 2016).

- the technology that lawyers use to run their practices . . .
- a traditional realm — understanding the technology used by our clients to design or manufacture products or to offer particular services
- the technology used to present information in the courtroom.

1.1.2 Widespread Adoption

In the eleven years since the change to the Model was implemented, forty states have adopted some form of this technology competence requirement for lawyers.[7] The vast majority of those

- Alaska (2017)
- Arizona (2015)
- Arkansas (2014)
- California (2021)
- Connecticut (2014)
- Delaware (2013)
- Hawaii (2022)
- Idaho (2014)
- Illinois (2016)
- Indiana (2018)
- Iowa (2015)
- Kansas (2014)
- Kentucky (2018
- Louisiana (2018)
- Massachusetts (2015)
- Michigan (2020)
- Minnesota (2015)
- Missouri (2017)
- Montana (2016)
- Nebraska (2017)
- New Mexico (2013)
- New York (2015)
- North Carolina (2014)
- North Dakota (2016)
- Ohio (2015)
- Oklahoma (2016)
- Pennsylvania (2013)
- Tennessee (2017)
- Texas (2019)
- Utah (2015)
- Vermont (2018)
- Virginia (2016)
- West Virginia (2015)
- Wisconsin (2017)
- Wyoming (2014)

A few states have made more noteworthy modifications or taken different approaches entirely.

1.1.3 Notable Variations

Colorado, Florida, New Hampshire, South Carolina, and Washington have each made some noteworthy modifications to the model comment in their implementations:

- Colorado made their version place a greater emphasis on communications technologies and protecting client data and communications.[8]
- Florida's version adds an explicit technology CLE requirement and explicitly addresses the role of technical experts in fulfilling the duty. [9]
- New Hampshire's variation adds qualifiers stressing reasonable efforts and evaluation against peers.[10]
- South Carolina's version limits the scope[11] from "relevant technology" to "technology the lawyer uses to provide services to clients or to store or transmit information related to the representation of a client."

[7]Robert Ambrogi, Tech Competence, LAWSITES, https://www.lawsitesblog.com/tech-competence (last visited July 2, 2021).
[8]Rule Change 2016(04) to Colo. Model Rules of Prof'l Conduct (2016), available at https://www.courts.state.co.us/userfiles/file/Court_Probation/Supreme_Court/Committees/Rules_of_Professional_Conduct_Committee/2016(04).pdf.
[9]*In re: Amends. to Rules Regulating the Florida Bar* 4-1.1 and 6-10.3, No. SC16-574 (Fla. Sept. 29, 2016), available at http://www.abajournal.com/files/OP-SC16-574_AMDS_FL_BAR_SEPT29_(1)_copy.pdf.
[10]*Order adopting amendments to court rules effective January 1, 2016* (N.H. Nov. 10, 2015), available at https://www.courts.state.nh.us/supreme/orders/11-10-15-Order.pdf.
[11]*Re: Amendments to Rules 1.0, 1.1, and 1.6, Rules of Professional Conduct, Rule 407, South Carolina Appellate Court Rules*, Case No. 2019-000318 (S.C. Nov. 27, 2019), available at https://www.sccourts.org/whatsnew/displayWhatsNew.cfm?indexId=2433.

- Washington's version adopted the model comment but also added an additional comment[12] about the potential role of that state's Limited License Legal Technicians[13]:

> In some circumstances, a lawyer can also provide adequate representation by enlisting the assistance of an LLLT of established competence, within the scope of the LLLT's license and consistent with the provisions of the LLLT RPC.

1.2 THE CALIFORNIA APPROACH

California did not formally adopt the model change until 2021,[14] but six years earlier, it took another approach to ensuring technology competence for eDiscovery. In 2015, it promulgated a detailed ethics opinion establishing a duty of technology competence for eDiscovery. Formal Opinion No. 2015-193[15] established that:

> Attorneys who handle litigation may not ignore the requirements and obligations of electronic discovery. Depending on the factual circumstances, **a lack of technological knowledge in handling e-discovery may render an attorney ethically incompetent to handle certain litigation matters involving e-discovery, absent curative assistance** . . . . [emphasis added]

This opinion went beyond just establishing a general duty, however. It also identified nine core competency requirements necessary to fulfill this duty of technology competence for eDiscovery:

1. "initially assess e-discovery needs and issues, if any"
2. "implement/cause to implement appropriate ESI preservation procedures"
3. "analyze and understand a client's ESI systems and storage"
4. "advise the client on available options for collection and preservation of ESI"
5. "identify custodians of potentially relevant ESI"
6. "engage in competent and meaningful meet and confer with opposing counsel concerning an e-discovery plan"
7. "perform data searches"
8. "collect responsive ESI in a manner that preserves the integrity of that ESI"
9. "produce responsive non-privileged ESI in a recognized and appropriate manner"

This list of requirements has been widely discussed as a useful model for all attorneys seeking to fulfill their duty of technology competence for eDiscovery.

[12]Wash. Model Rules of Prof'l Conduct R 1.1 (2021), available at https://www.courts.wa.gov/court_rules/pdf/RPC/GA_RPC_01_01_00.pdf.
[13]Robert Ambrogi, *Washington state moves around UPL, using legal technicians to help close the justice gap*, ABA JOURNAL, https://www.abajournal.com/magazine/article/washington_state_moves_around_upl_using_legal_technicians_to_help_close_the (Jan. 1, 2015).
[14]*Robert Ambrogi, California Becomes 39th State To Adopt Duty Of Technology Competence,* LAWSITES, https://www.lawsitesblog.com/2021/03/california-becomes-39th-state-to-adopt-duty-of-technology-competence.html (Mar. 24, 2021).
[15]The State Bar of California Standing Committee On Professional Responsibility and Conduct, *Formal Opinion No. 2015-193* (June 30, 2015), available at https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL 2015-193 %5B11-0004%5D (06-30-15) - FINAL.pdf.

1.2.1 Initially Assess eDiscovery Needs and Issues, if Any

The first requirement is that an attorney – or an attorney collaborating with an eDiscovery expert – be able to spot eDiscovery implications at the outset of each new matter. This requirement in some ways incorporates the other eight within it, as it asks you to think ahead about eDiscovery needs and issues that might arise throughout the course of the upcoming matter.

As several of the following requirements make clear, the most important things to be able to assess initially are (a) potential sources of ESI that will need to be considered and (b) any risks of loss associated with those sources that must be mitigated. Many kinds of mistakes can be remedied further down the road, but the loss of unique, relevant ESI cannot.

1.2.2 Implement/Cause to Implement Appropriate ESI Preservation Procedures

As we just noted, acting quickly to identify and prevent the loss of ESI sources is core to fulfilling the duty of eDiscovery competence. ESI spoliation remains a frequent issue – particularly in the gray area where new devices, applications, or services are transitioning from niche adoption to mainstream use. The first and most important step for preservation in most instances is the issuance of an effective legal hold, and the second is monitoring ongoing compliance with that hold (including the suspension of automatic janitorial functions). The third is moving quickly to collect and preserve a copy of any ESI source that is at too a high risk of loss or alteration to preserve in situ (e.g., smartphones, Slack channels).

1.2.3 Analyze and Understand a Client's ESI Systems and Storage

This requirement is the one most likely to require the assistance of technical experts, both your own and your client's. Every organization has a unique combination of enterprise, departmental, and individual computers, devices, and software (as well as third-party service providers and other potential sources). Moreover, each organization has its own standard operating procedures – both formal, documented ones and unofficial ones – that dictate how things are created, where things are stored, and for how long.

Untangling that unique mess to identify all the places that potentially-relevant ESI may be hiding typically requires the involvement of:

- Someone with intimate knowledge of those systems and practices (i.e., organization IT)
- Someone who understands the relevant legal scope and likely discovery obligations (i.e., in-house or outside counsel)
- Someone who can understand the technical details presented and assess them against the scope and obligations (i.e., an internal or external eDiscovery expert)

1.2.4 Advise the Client on Available Options for Collection and Preservation of ESI

There is obvious overlap between this requirement and the requirements above, but as we have noted, avoiding spoliation of ESI is at the heart of the duty of eDiscovery competence. This additional requirement is primarily aimed at making sure practitioners understand the range of data handling options available and the importance of maintaining forensic soundness.

For example, this would encompass understanding the importance of metadata, how easily it is altered, and how to ensure its preservation. This would also extend to understanding the risks associated with allowing self-collection, to understanding (at least at a high level) imaging and targeted collection options, and to considering newer remote collections solutions. As with the requirement above, this requirement is often fulfilled with the assistance of an expert that can provide a greater depth of both technical knowledge and collection experience.

1.2.5 Identify Custodians of Potentially Relevant ESI

This requirement fits hand-in-glove with the above requirements, which are more focused on the source systems and devices than the people wielding them. In addition to being able to identify and address those source systems and devices, practitioners need to be able to identify the key individual custodians. Since Zubulake V[16] in 2004, the phrase "key players" has been used to describe these essential custodians within an organization. Key players are those with direct knowledge of the underlying events or those most likely to have relevant information or materials, including ESI. They are often also the best source for information about how ESI is actually created, handled, shared, and stored on a day-to-day basis within the organization.

Beyond individual custodians, you must also be able to identify the custodians responsible for other kinds of potential ESI sources, such as those individuals who administer departmental or enterprise systems, those who oversee outsourced functions and outside services, and those who handle issuance and recycling of employer-issued laptops and mobile devices.

1.2.6 Engage in Competent and Meaningful Meet and Confer with Opposing Counsel Concerning an eDiscovery Plan

After all of the initial investigative and scoping steps, and after initial preservation is assured, the next requirement an attorney must be prepared to fulfill is engaging in meaningful discussion about eDiscovery during the meet and confer with opposing counsel. Fulfilling the initial five requirements is a condition precedent to being able to fulfill this one.

Negotiating meaningfully about an eDiscovery plan requires already having some concrete knowledge of what ESI exists, where it exists, and in what forms it exists. It requires having already considered the applicable collection options and their associated limitations, risks, and costs, as well as any ESI that may not be reasonably accessible due to burden or cost. Additionally, it requires

[16] *Zubulake v. UBS Warburg LLC*, 229 F.R.D. 422 (S.D.N.Y. 2004), available at https://casetext.com/case/zubulake-v-ubs-warburg-llc-3.

looking ahead to the later steps in the discovery process (and the later requirements in this list) to assess potential search protocols, review methodologies, and production plans.

It is entirely too common for parties to commit themselves to eDiscovery plans that wind up being either excessively burdensome or technically impossible in some way, because they negotiated the plans without adequate knowledge of the actual facts on the ground or without adequate understanding of (or expert guidance about) the technical realities associated with later steps. On the other hand, when handled effectively, negotiation of an eDiscovery plan can provide an opportunity to dramatically limit the time and cost of discovery through agreed limitations on scope, through preemption of downstream conflicts, or through adoption of a phased discovery plan.

1.2.7 Perform Data Searches

Once actual discovery work has begun, the next requirement attorneys must be able to satisfy – either on their own or with the assistance of an appropriate expert – is the effective execution of data searches. This applies both to searches of source systems for materials to collect and to searches of processed materials for the right materials to review and produce. Searching effectively at any point in the eDiscovery process requires understanding both substantive and technical realities:

Substantively, you must have some understanding of the content of the source materials and the likely content of the specific materials you are seeking within them. You must have some sense of the language used generally and some idea where the specific language you seek might be found.

Technically, you must have some understanding of the capabilities and limitations of the specific search tools you are using. For example, some tools search automatically within nested content (e.g. attachments and container files), and some tools cannot do so at all. Some tools can understand complex Boolean logic, some can only handle simple keywords, and others have their own custom search syntax that must be followed. Some have limitations on where they can search or how many results they can return.

Failure to understand these realities increases the chances of ineffective searches and of difficult-to-detect gaps in your results.

1.2.8 Collect Responsive ESI in a Manner that Preserves the Integrity of that ESI

We touched a bit on the goal of this requirement above, in our discussion of the requirement that attorneys be able to advise their clients on options for preservation and collection of ESI. Unlike physical documents, electronic documents are very easily changed – even accidentally. They can be changed by being moved or copied or forwarded. They can be changed simply by being opened and viewed. Consequently, ESI must be handled very carefully to collect it and work with it in a way that both preserves the original and produces accurate copies for use in a legal matter.

Avoiding self-collection strategies and informal data handling practices is essential to fulfilling this aspect of the duty of technology competence for eDiscovery. Metadata must be preserved, forensic soundness must be ensured, and chain of custody must be documented.

1.2.9 Produce Responsive Non-Privileged ESI in a Recognized and Appropriate Manner

The final requirement is that attorneys be able to produce ESI effectively. This is another requirement that is almost always fulfilled in collaboration with an internal expert or an external service provider, but it is important for attorneys to understand the range of possibilities and their differing requirements, limitations, and costs.

Depending on what is negotiated or required, ESI production may be as simple as creating a few PDF files, or as complicated as custom load files with extracted text and redacted, Bates-numbered page images. Relational database sources will also require negotiations about what reports or exports to generate and how best to present that information.

How materials are produced affects how long they take to prepare and how easily they can be searched, reviewed, and used later in depositions and at trial. Negotiating production format, including details like whether and what metadata will be provided, can both ensure maximum usability of what you receive and preempt disputes over what you produce and how you produce it. Failure to understand and negotiate in advance remains a common cause of discovery disputes.

1.3 KEY TAKEAWAYS

There are three key takeaways from this chapter to remember:

1. In 2012, the ABA promulgated a change to its Model Rules of Professional Conduct making "the benefits and risks associated with relevant technology" a required subject for maintaining competence.
2. Since then, thirty-nine states have adopted that change or a variation on it, including California, which also issued Formal Opinion No. 2015-193 identifying nine core competencies required to fulfill the duty of technology competence for eDiscovery.
3. Those nine core competencies provide a useful model for anyone seeking to ensure their own technology competence for eDiscovery, emphasizing effective identification and preservation of ESI, effective collection and production of that ESI, and effective negotiation about those processes.

Chapter 2

In the Beginning: Identification and Preservation Fundamentals

About this Chapter

In this chapter, we will discuss fundamentals of identification and preservation that all practitioners should know, including: the legal and technological scope of the duty, imagining the possibilities, investigating the realities, the role of legal holds, and common pitfalls and key takeaways.

2.1 IN THE BEGINNING

Identification and preservation are the first and most fundamental phases of an electronic discovery effort. The duty of (identification and) preservation is a foundational concept in our legal system that grows out of the common law concept[1] of "spoliation," which is nearly 300 years old.[2] Essentially, if courts exist to make determinations about disputed facts, and if the trier of fact must make those determinations using the available evidence, then no litigant should be allowed to gain advantage in those determinations by hiding or destroying relevant evidence before the trier of fact can consider it.

As we will see in numerous contexts, ESI spoliation remains a frequent issue – particularly in the gray area where new devices, applications, or services are transitioning from niche adoption to mainstream use. Hence the importance of these phases in an eDiscovery effort: almost every other type of failure can be fixed with adequate time and money, but once unique, relevant ESI is gone, it's gone.

2.1.1 Identification, Preservation, and the Duty of Competence

Beyond simply being important, the ability to successfully identify and preserve relevant ESI may also be an ethical requirement for attorneys to fulfill their duty of technology competence. For example, the California duty of technology competence for eDiscovery,[3] explicitly discusses identification and preservation skills in four of its nine core requirements:

- "initially assess e-discovery needs and issues, if any"
- "implement/cause to implement appropriate ESI preservation procedures"
- "advise the client on available options for collection and preservation of ESI"
- "identify custodians of potentially relevant ESI"

In this conception of it, avoiding spoliation of ESI is at the heart of the duty of eDiscovery competence.

2.1.2 Triggers for the Duty of (Identification and) Preservation

The duty to identify and preserve documents often arises even before a case is actually filed or commenced, because the duty arises not when there is litigation but when there is reasonable anticipation of litigation (or agency action, etc.). As explained in "Guideline 1" of The Sedona Conference Commentary on Legal Holds, Second Edition: The Trigger & The Process[4]:

> A reasonable anticipation of litigation arises when an organization is on notice of a

[1]Margaret M Koesel & Tracey L Turnbull, *Spoliation of Evidence: SANCTIONS AND REMEDIES FOR DESTRUCTION OF EVIDENCE IN CIVIL LITIGATION* 2 (3rd Ed. 2013), available at https://www.americanbar.org/content/dam/aba-cms-dotorg/products/inv/book/214612/Chapter%201.pdf.
[2]*Id.* at xv, available at https://www.americanbar.org/content/dam/aba-cms-dotorg/products/inv/book/214612/Introduction.pdf; see also *Armory v. Delamirie*, 93 Eng. Rep. 664 (K.B. 1722), available at https://www.bailii.org/ew/cases/EWHC/1722/J94.html.
[3]The State Bar of California Standing Committee on Professional Responsibility and Conduct, *Formal Opinion No.* 2015-193 (2015), available at https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL 2015-193 %5B11-0004%5D (06-30-15) - FINAL.pdf.
[4]The Sedona Conference, *Commentary on Legal Holds, Second Edition: The Trigger & The Process*, 20 SEDONA CONF. J. 341 (2019), available at https://thesedonaconference.org/publication/Commentary_on_Legal_Holds.

> credible probability that it will become involved in litigation, seriously contemplates initiating litigation, or when it takes specific actions to commence litigation.

Examples of triggering events include: discovery of a legal or regulatory violation by an employee, receipt of a legal hold notice from a regulatory agency, hearing a terminated employee threaten suit, receipt of an actual complaint or subpoena, and many more.

2.2 LEGAL AND TECHNOLOGICAL SCOPE

The first thing you must know to undertake effective identification and preservation activities is the potential legal and technological scope of things for which you might need to be looking. For our purposes, that scope comes from the Federal Rules of Evidence and Civil Procedure and relevant case law, and it can be quite broad.

2.2.1 The Scope of the Duty of (Identification and) Preservation

The scope of potential discovery – and, therefore, of the duty to (identify and) preserve – is deliberately broad, which is consistent with our court system's emphasis on truth-seeking over gamesmanship. As stated in one court decision[5] involving discovery sanctions:

> Litigation is not a game. It is the time-honored method of seeking the truth, finding the truth, and doing justice. When a corporation and its counsel refuse to produce directly relevant information an opposing party is entitled to receive, they have abandoned these basic principles in favor of their own interests.

In its simplest form, the potential scope of discovery – and, therefore, preservation – for ESI has four elements:

1. Documents
2. In your possession, custody, or control
3. That are potentially relevant
4. That are unique

Documents

The definition of "documents" provided by the Federal Rules of Civil Procedure is expansive enough to encompass almost any sort of material in any format. Rule 34(a)(1)(A)[6] states that it covers "documents and 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 . . . [emphasis added]

The Committee Notes[7] on the rule emphasize that this is intended "to be broad enough to cover The first thing you must know to undertake effective identification and preservation activities is the potential legal and technological scope of things for which you might need to be looking. For our purposes, that scope comes from the Federal Rules of Evidence and Civil Procedure and relevant case law, and it can be quite broad.

[5] *Haeger v. Goodyear Tire & Rubber Co.*, 813 F.3d 1233, 1237 n.1 (9th Cir. 2016), available at https://scholar.google.com/scholar_case?case=272820219719515977 5.
[6] Fed. R. Civ. P. 34(a)(1)(A), available at https://www.law.cornell.edu/rules/frcp/rule_34.
[7] Fed. R. Civ. P. 34 advisory committee's note, available at https://www.law.cornell.edu/rules/frcp/rule_34.

Possession, Custody, or Control

Rule 34(a)(1)[8] also specifies that the scope of discovery and preservation extends to those documents within "the responding party's possession, custody, or control." This phrase means that you are responsible, not just for the materials you physically or electronically possess, but for any that you legally control. Materials maintained by third parties on your behalf are treated the same way as the records you actually possess yourself. If you have the right (or, in some cases, the ability) to obtain it, you are responsible for preserving and producing it.

Please note that there are actually three distinct standards for how far "possession, custody, or control" is deemed to extend, depending on your jurisdiction: "Legal Right," "Legal Right Plus Notification," and "Practical Ability." More information is available in The Sedona Conference Commentary on Rule 34 and Rule 45 Possession, Custody, Or Control.[9]

Potentially Relevant

Among the "documents" that are in your "possession, custody, or control," the ones that may be discovered and must be preserved are those that are relevant. Relevance is defined broadly by Federal Rule of Evidence 401.[10] That rule dictates that evidence is relevant if "it has any tendency to make a fact more or less probable than it would be without the evidence" and "the fact is of consequence in determining the action." The Committee Notes[11] to the rule state explicitly that this is an intentionally low bar because "[a]ny more stringent requirement is unworkable and unrealistic." Thus, any documents in your possession, custody, or control that have any tendency to make any fact of consequence more or less likely are relevant, potentially discoverable, and required to be preserved.

Unique

Finally, the scope of potential discovery and required preservation is limited to materials meeting the above criteria that are also unique. As specified by Rule 26(b)(2)(C),[12] discovery is not meant to be "unreasonably cumulative or duplicative." For ESI in particular, this is important, as it is in the nature of electronic systems to create numerous identical copies of materials, both for operation and for backup. Generally, there will be no additional evidentiary value to preserving numerous identical copies of the same materials.

Other Limitations

Beyond those four elements, there are two additional potential limitations on the scope of discovery that are less relevant to the question of preservation scope:

First, as specified in Rule 26(b)(1),[13] the scope of discovery is limited to that which is "proportional to the needs of the case." Because any disputes over proportionality cannot be identified and resolved by the court until the matter is already underway, parties should not be quick to assume disproportionality and skip preservation.

Second, as specified in Rule 26(b)(2)(B),[14] "[a] party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because

[8]Fed. R. Civ. P. 34(a)(1), available at https://www.law.cornell.edu/rules/frcp/rule_34
[9]The Sedona Conference, *The Sedona Conference Commentary on Rule 34 and Rule 45 "Possession, Custody, or Control,"* 17 SEDONA CONF. J. 468, 482 (2016), available at https://thesedonaconference.org/publication/Commentary_on_Rule_34_and_Rule_45_Possession_Custody_or_Control.
[10]Fed. R. Evid. 401, available at https://www.law.cornell.edu/rules/fre/rule_401.
[11]Fed. R. Evid. 401 advisory committee's note, available at https://www.law.cornell.edu/rules/fre/rule_401.
[12]Fed R. Civ. P. 26(b)(2)(C), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[13]Fed. R. Civ. P. 26(b)(1), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[14]Fed. R. Civ. P. 26(b)(2)(B), available at https://www.law.cornell.edu/rules/frcp/rule_26.

of undue burden or cost." This is another type of proportionality requirement, and as with the general proportionality requirement, any disputes over it cannot be identified and resolved by the court until the matter is already underway.

Preservation can always be stopped if it's later determined to be unnecessary, but lost unique materials can never recovered if they're later determined to have been necessary after all.

2.2.2 Ongoing Source Evolution

We noted above that the rules were written to accommodate the ongoing evolution of computer technology and ESI sources. It's worth emphasizing here that this means the potential technological scope of the duty of (identification and) preservation is a moving target, evolving as the technology and services that produce relevant ESI do. For example:

- In 2017, Microsoft rolled out a new chat and collaboration application called Teams to compete with Slack.
- Despite Slack's four-year head start, by November 2019, Microsoft Teams had surpassed Slack with 20 million daily users.[15]
- This rapid growth was turbocharged by the pandemic and consequent shift to remote and hybrid work, which resulted in geometric growth for Microsoft Teams.
- Daily active users tripled in 2020, and then they more than doubled again in 2021. By the end of 2022, Teams had over 270 million monthly active users.[16]

Five years ago, Teams didn't exist, and three years ago, it was only just becoming a source you were expected to consider. Today, it's as expected as email. Technology will continue to leap forward, and legal expectations will follow slowly behind.

2.3 IMAGINING THE POSSIBILITIES

Now that we have reviewed the importance of performing effective identification and preservation and discussed the potential legal and technological scope of what we might need to identify and preserve, we are ready to start discussing the actual identification process. We're going to break down the identification process into two parts: imagination and investigation.

[15]Mary Jo Foley, "Microsoft says it has 20 million daily active Teams users," ZDNET (Nov. 19, 2019), available at https://www.zdnet.com/article/microsoft-says-it-has-20-million-daily-active-teams-users/.
[16]Lionel Sujay Vailshery, "Microsoft Teams: number of daily active users 2019-2022," Statista (Jan 13, 2023), available at https://www.statista.com/statistics/1033742/worldwide-microsoft-teams-daily-and-monthly-users/.

2.3.1 A World of Imagination

We don't spend a lot of time talking about imagination in legal practice, but it's pretty essential to effective identification. Each new discovery project can be fairly opaque at the outset. You may know the main issue and underlying event, but you may not know much beyond that about the relevant individuals, the various potential claims and defenses, etc. All of that knowledge will only be developed as you proceed with the project.

The first step, then, must be brainstorming to figure out what and who might be relevant. Doing this effectively requires collaborating with:

- Individuals with direct knowledge of the relevant legal issues
- Individuals with direct knowledge of the relevant factual issues
- Individuals with direct knowledge of the organization's IT systems and devices

Essentially, in-house counsel, outside counsel, internal IT, any external collection resources, and any relevant senior employees all need to contribute to this effort. Starting with what you know about the type of matter, the underlying facts, and the involved individuals, you must extrapolate what types of relevant materials are likely to exist within the organization and where (or in whose custody) those materials are likely to be.

This process can be aided by checklists of potential sources (like those used for custodian surveys/interviews), but it is a fundamentally imaginative exercise: imagine the events at issue in the context of normal organizational operations and think about what might have been generated. For example:

- Might there be departmental records, like HR files?
- Could there be useful data about the events in your ERP systems?
- Would employees have discussed the events via an internal chat client?
- Maybe the relevant office used shared network folders?
- Perhaps copies of deleted records exist on back-up tapes?

Additionally, it is beneficial to imagine what distinctive characteristics relevant materials from these sources might bear, i.e. how you would try to find them if running searches for them:

- Are you seeking evidence of intent in communications between certain employees?
- Are you looking for evidence of internal awareness in executive meeting minutes?
- Will relevant documents contain certain keywords, like a name or project code?
- Are you looking for contracts executed with a particular party or on certain dates?
- Are you looking for metadata evidence of an employee altering key documents?

Having some ideas about distinctive characteristics of this type will also be helpful to you as you move on to investigation and preservation.

2.3.2 Finding Your Key Players

A key part of this exercise is figuring out who your key players are likely to be for the matter. Key players are those with the direct knowledge of the underlying events and those most likely to have relevant materials. The phrase rose to prominence after Zubulake V[17] in 2004, which used it to describe the essential recipients of a legal hold within an organization.

In addition to the individual employees, managers, and executives involved in the events underlying the matter, it is also critical not to forget other types of important players that may be in possession or control of relevant materials:

- **Individuals responsible for enterprise or departmental IT systems** – organizations may have any number of enterprise systems (e.g. email, backup, or document management) and departmental systems (e.g. benefits, payroll, research, or compliance) that may contain relevant information beyond that held directly by individual custodians. The individuals responsible for managing such systems are important players to include in your identification and preservation process to ensure materials aren't missed or lost.
- **Third-party service providers** – organizations very commonly outsource one or more business functions, like payroll or benefits (or even email), to specialized third-party service providers, and the data they possess on your organization's behalf may contain relevant materials you need to identify and preserve. These service providers (or the relevant individuals within them) are also important players to include in your identification and preservation process to ensure materials aren't missed or lost.

Completing your initial brainstorming by making a list of all your expected key players – including important enterprise, departmental, or third party players – will help prepare you for the investigation and preservation steps that come next.

2.4 INVESTIGATING THE REALITIES

A variety of investigative options are available for finding out how reality lines up with the brainstorming you've done to get started. The most important are: targeted interviewing, data mapping, and sampling. Which one(s) will be most useful to you will depend on your specific project – in particular, how large and diverse your brainstorming has led you to believe your project will be. For example:

- The larger your project, the more investigative steps you'll need to take
- The more systems and sources by count, the more useful a data map is
- The more custodians by count, the more useful sampling is

We will discuss each of the three primary options in turn.

[17] *Zubulake v. UBS Warburg LLC*, 229 F.R.D. 422 (S.D.N.Y. 2004), available at https://casetext.com/case/zubulake-v-ubs-warburg-llc-3.

2.4.1 Targeted Interviews

Targeted interviews are the easiest investigative step and a common first one. In this context, conducting targeted interviews is like conducting a limited number of custodian interviews with some important and key players. This process is typically less formal (i.e., no full script) and less complete (i.e., not all individual custodians are included) than a full custodian interview process, which might come later in the project.

Your goal in the targeted interviews is to review your brainstormed lists of materials and people with individuals that have some direct knowledge of what likely exists, where it would be, and who else might know things or possess relevant materials. This would include talking to individuals with knowledge of any potentially relevant enterprise or departmental systems, as well as any relevant third party service providers.

2.4.2 Data Mapping

Your next investigative option is data mapping. Data mapping is the process of mapping the various data stores and sources in an organization. Many organizations do some version of this already for non-legal purposes. For example, the IT or IS department may have maps of the organization's servers, computers, and enterprise systems, along with directories of installed software. Ideally, data mapping for legal activities would be undertaken on a proactive, organization-wide basis rather than in response to a specific matter, but engaging in some targeted, reactive data mapping is better than none and well worth doing.

In this context, you would be working your way down your potential materials/hypothetical sources list, reviewing them with relevant individuals (from IT/IS, Records Management, etc.) and reviewing relevant documentation, attempting to flesh out that list with concrete details. What you will be attempting to build is less a literal map than a spreadsheet or matrix. Your final product will be a searchable, sortable, filterable reference tool (e.g., a spreadsheet listing sources in rows and relevant details about them in columns). Important things to note about each source during this sort of reactive data mapping include:

- Owner/manager of source (e.g., specific IT contact, department manager, or custodian)
- Desired materials expected to be there (including expected formats, dates, etc.)
- Expected volume of materials from source (e.g., record count, file volume)

- Available native search and export tools/features, if any, and relevant details
- Risk to those materials from automated janitorial functions or other normal processes

Gathering and organizing this information (and additional details, as time and circumstances permit) will equip you to better plan your needed preservation (and later collection) activities. Plus, from a data targeting perspective, the more information available to you about what you have and where it is, the more narrowly and accurately you can target what gets preserved and collected for subsequent phases of project work.

Sampling

The final investigative option we'll discuss is sampling. In the context of eDiscovery, sampling is used to refer to both judgmental and statistical sampling. Judgmental sampling is the informal process of looking at parts of something large to get an anecdotal sense of the whole, while statistical sampling refers to formal sampling to take a defined measurement. In the identification and preservation phases at the beginning of a matter, you will primarily be engaged in judgmental sampling.

Judgmental sampling is essentially what you're doing when you select key individuals for targeted interviews, using them as proxies for the whole list of hypothetical custodians. More importantly, though, judgmental sampling is the way you will learn about what's actually on various sources and systems that, unlike custodians, cannot self-report to you. This kind of judgmental sampling might take a variety of forms, such as:

- Searching electronic mailboxes to test for relevance
- Indexing some backup tapes to test for unique materials in backups
- Collecting representative custodians' devices to check for relevant materials

These efforts will be aided by the time you spent brainstorming potential distinctive characteristics of the materials you are seeking.

Depending on your project's scale and timeline, you may end up proceeding from judgmental sampling at the beginning to formal statistical sampling before collection (when it may be worth the effort and expense to get some firm measurements for decision-making and process negotiations). This has become especially true in this era of increased focus on proportionality, but those are generally questions to address after effective identification and preservation have already taken place (so that nothing unique is lost in the meantime).

2.5 THE ROLE OF LEGAL HOLDS

Once you've completed your imagination and investigation activities and have identified the potentially-relevant materials within your organization, you are ready to take steps to actually preserve those materials. The first and most important aof those steps is the issuance of a legal hold instructing the custodians of potentially-relevant materials regarding the need to preserve them.

It's true that legal holds do not preserve data themselves, but they are the critical first step in the preservation process, ensuring that materials survive in situ long enough for you and your team to go get them. You are literally saying to everyone – just as Sam & Dave sang in 1966: "Hold On, I'm Coming."[18]

2.5.1 Legal Holds and eDiscovery

Formal, written legal holds became the focus of much attention in eDiscovery after the Zubulake V[19] ruling in 2004, in which a party was sanctioned for failing to issue a hold or take other necessary steps to ensure the preservation of relevant materials. In subsequent years,[20] this decision was cited in numerous others, and written legal holds became central to an effective eDiscovery preservation process.

For a time, the failure to issue a written legal hold was treated as per se gross negligence.[21] That absolute requirement for a hold in writing was softened by subsequent cases,[22] however, which allowed for the possibility of circumstances in which oral holds or other approaches to preservation may be appropriate.

2.5.2 Issuing a Legal Hold

When preparing a legal hold for issuance, there are five essential elements that should be included. An effective legal hold should include information regarding:

1. **The legal obligations associated with the hold** – This should include some explanation of the duty to preserve, the legal consequences for the organization if it is not fulfilled, and any internal consequences for employees who violate it. It is often helpful to point out that a request for individuals to preserve materials is a common legal step and is not an indication recipients are in any trouble.
2. **The substantive scope of what must be preserved** – This may include describing the underlying events, the relevant individuals inside and outside the organization with whom communication may have taken place, and more. This should also include the applicable time range, if any, and whether the hold applies going forward to newly created materials as well.
3. **The types of materials that must be preserved** – This should include lists of relevant devices (e.g., laptops, phones, thumb drives), of file types (e.g., email, spreadsheets, text messages), and of example documents (e.g., internal financial reports, contract negotiation messages, annotated contract drafts).

[18]Sam & Dave, "Hold On, I'm Coming" (1966), available at https://www.youtube.com/watch?v=Fowldx4hRtI.
[19]*Zubulake*, 229 F.R.D. 422.
[20]Victor Li, *Looking back on Zubulake, 10 years later*, ABA JOURNAL, http://www.abajournal.com/magazine/article/looking_back_on_zubulake_10_years_later (Sept. 1, 2014).
[21]Rachel S. Fendell, *Impact Of Chin Decision On Pension Committee*, MONDAQ, https://www.mondaq.com/unitedstates/disclosure-electronic-discovery-privilege/190306/impact-of-chin-decision-on-pension-committee (Aug. 6, 2012).
[22]*Chin v. Port Auth. of N.Y. & N.J.*, 685 F. 3d 135 (2nd. Cir. July 10, 2012), available at https://scholar.google.com/scholar_case?case=11269039069845908318.

4. **The process that is to be used for preservation and collection** – These are the specific instructions the recipients of the hold are to follow for handling the materials they possess that are subject to the hold. Should they just preserve them in place? Take other steps? When and how will they be contacted about collection of those materials?
5. **How and with whom recipients may communicate about the hold** – This should include both any prohibitions on communication about the hold or the underlying matter with peers (to protect privilege), as well as instructions for who should be contacted with any questions about scope or process.

When drafting, it is important to remember that the hold must cover not only the devices and materials of individual custodians, but also relevant departmental and enterprise systems and any automated janitorial functions that may be running on them. Additionally, you may consider including: compliance confirmations to be returned, custodian surveys to aid in collection planning, or frequently asked questions to help recipients understand what's required.

Once you've created your hold, you are ready to issue it to the lists of key players, important system managers, and other custodians that you identified during your brainstorming and investigation activities.

2.5.3 Monitoring a Legal Hold

Beyond just issuing a legal hold, it is crucially important that you engage in some form(s) of ongoing compliance monitoring after hold issuance. As has been made clear in case[23] after case,[24] failure to check if individuals are actually complying, or to remind them as needed, can be just as consequential as failing to issue the hold in the first place.

Common steps taken to ensure ongoing compliance with a legal hold include:

6. **Employee verification** – having employees sign a document or electronic form, or send an email, confirming that they have received the hold, understood the hold, and will comply with the hold; this is typically covered as part of the hold itself
7. **System verification** – just as individual custodians must confirm their understanding and compliance commitment in writing, so those responsible for suspending janitorial functions on enterprise or departmental systems can be required to do the same
8. **Recycling verification** – the same requirement can be applied to the individuals responsible for the recycling of backup tapes and employee devices
9. **Spot checking** – it also advisable to establish a regular schedule for checking in with at least a sampling of the subject custodians (checking everyone may not be feasible) to check that they are in fact complying and materials are being preserved
10. **Reissuance** – since legal matters and the holds associated with them can continue for months or years, it is also advisable to establish a schedule for periodic reissuance of the hold as a reminder to those it covers (quarterly is common); the specific scope of the hold may also need to be revised as a legal matter evolves and more is learned

As we noted above, a legal hold is not itself preservation, and if it is not followed by steps to ensure actual preservation takes place – like ongoing compliance monitoring, then whether or not a hold was issued doesn't really matter.

[23] *Pension Comm. of the Univ. of Montreal Pension Plan v. Banc of America Sec., LLC*, 685 F. Supp. 2d 456 (S.D.N.Y. Jan 15, 2010), available at https://www.courtlistener.com/opinion/1881971/univ-of-montreal-pension-plan-v-banc-of-am-sec/.
[24] *Chin*, 685 F. 3d 135.

2.6 COMMON PITFALLS

We have now reviewed how important identification and preservation are, how broad their scope might be, how to go about brainstorming and investigating to identify what needs to be preserved, and how to issue and monitor compliance with legal holds. What remains is to think about common preservation pitfalls, situations in which iammediate collection is called for, and investigations in which some custodians may also be bad actors with a reason to spoliate.

2.6.1 Common Pitfalls

The three most common pitfalls in identification and preservation all relate to technological blind spots:

1. **Failure to deactivate automatic janitorial functions** – many of the devices and systems in use within your organization will have some sort of automatic janitorial function in place, a feature that automatically deletes older materials to prevent the accretion of infinite emails, texts, etc. Failure to deactivate such features (or to collect before they delete relevant ESI) frequently results in the inadvertent loss of unique, relevant materials.
2. **Failure to account for third-party service provider materials** – as we noted above, third-party service providers may be storing relevant ESI belonging to your organization (potentially with their own storage limits or automated janitorial functions). Failure to account for such sources can result in the inadvertent loss or alteration of unique, relevant materials.
3. **Failure to consider newer technology and source types** – as we also noted above, new devices, applications, and services are appearing all the time, and custodian behavioral patterns are constantly changing to incorporate them. Failure to consider newer technology an source types can result in the inadvertent loss or alteration of unique, relevant materials.

These pitfalls can be avoided – in part – by making sure you consider all source types and service providers when doing your identification, by making sure all the custodians and service providers responsible for the relevant systems and services receive your legal hold notice, and by following up to make sure those recipients have actually deactivated or modified automatic janitorial functions where needed.

2.6.2 Immediate Collection

The other part of avoiding these pitfalls is proceeding promptly to actual collection for any source or source type you fear is at a high risk of loss or alteration. For example, smartphones entail a high risk of data loss due to the triple threats of automatic deletion of old messages, user deletion of messages, and frequent device replacements/upgrades. Other dynamic sources with an elevated risk of loss or alteration include ephemeral messaging applications, social media sources, and cloud-based collaboration tools. Preservation through collection ensures at-risk data is preserved for later analysis, review, and production.

2.6.3 Considerations for Investigations

Another scenario worth discussing is an investigation in which there is a possibility of bad actors among the relevant custodians. In such situations, there is a possibility that receiving a legal hold notice will prompt the bad actor(s) to intentionally destroy or alter materials prior to collection, spoliating evidence to hide their individual wrongdoing.

In matters where this is a concern, it may be necessary to undertake certain collection activities immediately – even before issuing a hold notice to all relevant custodians. By moving immediately to collect what you believe to be the key sources from the key custodians, you can prevent loss due to intentional spoliation, and preliminary analysis and review can begin while additional identification, preservation, and collection efforts are still ongoing in parallel.

There are several collection strategies that can be employed to acquire key data without alerting suspected employees. For example:

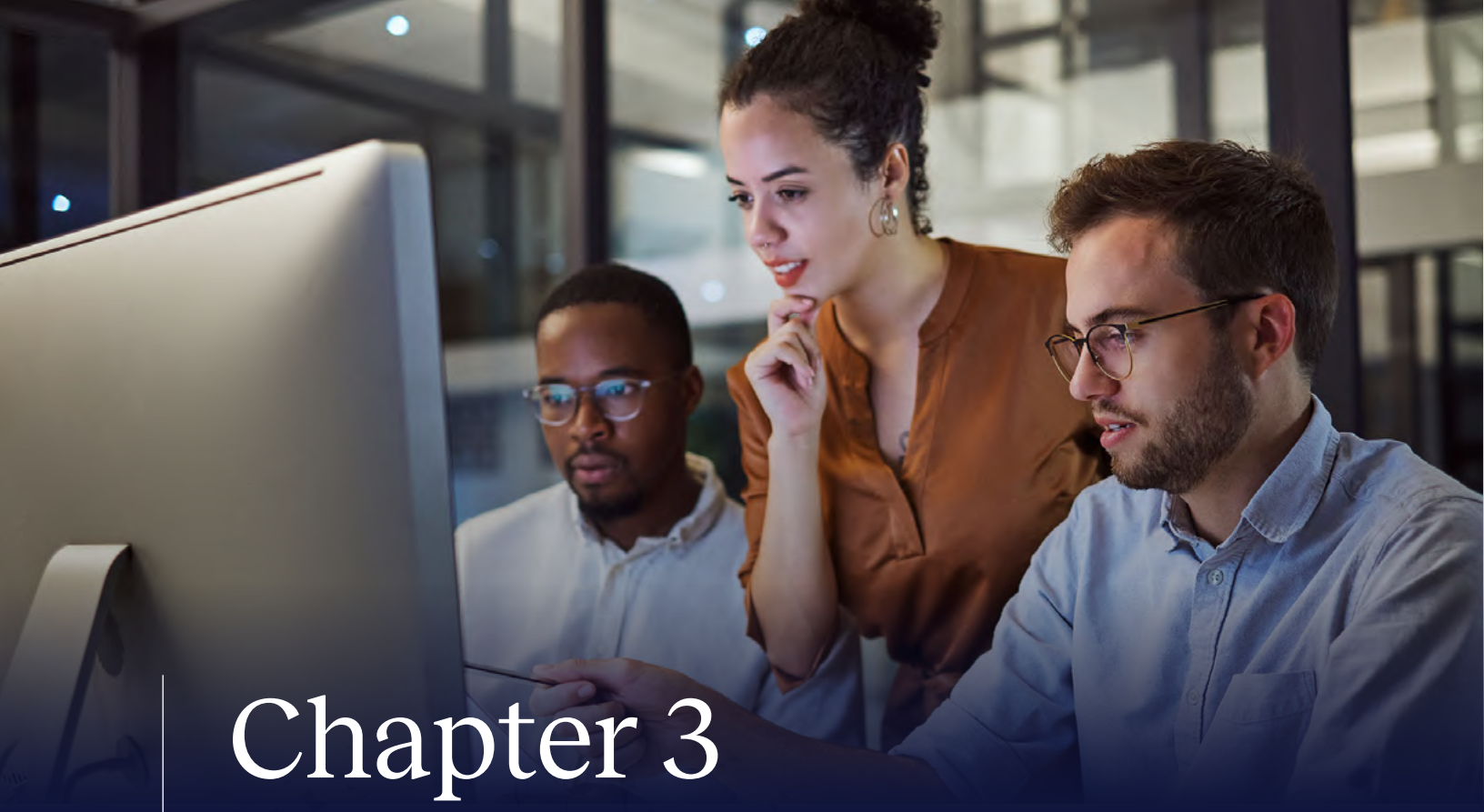
- IT can typically collect from employees' active corporate accounts for email, messaging, documents, etc. without alerting the employees, and IT can also take steps to preserve existing backups of those sources as needed
- For an individual, a laptop or computer (or mobile device) upgrade can be triggered, allowing IT to collect the current machine and image it without raising suspicions
- For a team or department, IT can require all laptops be left at desks overnight for required security or software updates, and images can be made during those hours

Because, in this context, speed and secrecy are most important, it is generally best to err on the side of over-collection (e.g., capturing full images, mailboxes, etc.) from those key sources to avoid missing something that could then be spoliated after hold issuance.

2.7 KEY TAKEAWAYS

There are five key takeaways from this chapter to remember:

1. Identification and preservation are at the very heart of attorneys' duty of technology competence for eDiscovery and are essential for preventing spoliation; almost every other type of failure can be fixed, but once unique, relevant ESI is gone, it's gone.
2. The legal scope of identification and preservation extends to all unique, potentially-relevant documents in your possession, custody, or control, and the technological scope may extend to any device or application, including smartphones, social media, and more.
3. Effective identification begins with brainstorming what materials might exist, where they might be, and what key players might have them or know more about them; then comes investigating the reality as needed through interviews, sampling, and data mapping.
4. Effective preservation begins with issuing a legal hold to the relevant individual and institutional custodians; the hold should address the legal obligations, the substantive scope, the types of materials, the process to be used, and the communication rules.
5. Issuance of a legal hold must be followed by ongoing monitoring of compliance with the hold, including written verifications, sampling, and periodic reissuance; it must also be followed (or accompanied) by efforts to promptly collect materials at a high risk of loss.



Chapter 3

The Grand Scavenger Hunt: Collection Fundamentals

About this Chapter

In this chapter, we will discuss collection fundamentals including: the scope of collection; how data is stored and recovered; the importance of metadata, forensic soundness and chain of custody; the risks of self-collection; other available collection approaches; and major source categories to be considered.

3.1 THE GRAND SCAVENGER HUNT

Since electronically-stored information (ESI) has become the norm in discovery, competence with technology has become an essential part of being an effective legal practitioner. With source types multiplying – including challenging sources like smartphones, social media, and collaboration tools, it is more important than ever for legal practitioners of all types to familiarize themselves with the fundamentals of collection so that they can assist in spotting potential issues and identifying appropriate solutions.

3.1.1 Collection and the Duty of Competence

Understanding the fundamentals of collection is also essential to fulfilling a lawyer's duty of technology competence, which exists in some form in forty states.[1] For example, as articulated in California's Formal Opinion No. 2015-193,[2] there are nine core requirements that lawyers must satisfy to fulfill their duty of technology competence for eDiscovery, two of which explicitly discuss collection: "advise the client on available options for collection and preservation of ESI" and "collect responsive ESI in a manner that preserves the integrity of that ESI." Another four of those nine requirements also necessitate an understanding of collection for their fulfillment ("initially assess e-discovery needs and issues, if any," "implement/cause to implement appropriate ESI preservation procedures," "analyze and understand a client's ESI systems and storage," and "identify custodians of potentially relevant ESI").

Thus, understanding the fundamentals of collection is essential to fulfilling a lawyer's duty of technology competence for eDiscovery in California and, likely, in many other states as well.

3.2 THE BROAD SCOPE OF COLLECTION

3.2.1 The Legal Scope of Collection

The practical scope of ESI collection is determined both by the actual requests from other parties and by your own information needs related to the matter. The maximum-possible scope is established by the Federal Rules of Civil Procedure (FRCP) or your state's equivalent ruleset. The FRCP establishes that scope as encompassing:

- Any documents or electronically-stored information
- In your possession, custody, or control
- That are relevant
- That are unique
- That are not unreasonably inaccessible because of undue burden or cost
- That are not disproportionate to the needs of the case

The first three criteria set a very broad potential scope for discovery collection. The definition of "documents or electronically stored information" provided by FRCP 34 and its accompanying committee notes[3] is expansive enough to encompass almost any sort of material in any format.

[1]Robert Ambrogi, *Tech Competence, LAWSITES*, https://www.lawsitesblog.com/tech-competence (last visited July 2, 2021).
[2]The State Bar of California Standing Committee On Professional Responsibility and Conduct, *Formal Opinion No. 2015-193* (June 30, 2015), available at https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL 2015-193 %5B11-0004%5D (06-30-15) - FINAL.pdf.
[3]Fed. R. Civ. P. 34, available at https://www.law.cornell.edu/rules/frcp/rule_34; Fed. R. Civ. P. 34 advisory committee's note, available at https://www.law.cornell.edu/rules/frcp/rule_34.

"[P]ossession, custody, or control" means that you are responsible, not just for the materials you physically or electronically possess, but for any that you legally control (or, potentially, that you have the practical ability to obtain). "Relevant" is also defined broadly, by Federal Rule of Evidence 401,[4] which states that evidence is relevant if "it has any tendency to make a fact more or less probable than it would be without the evidence" and "the fact is of consequence in determining the action."

The last three criteria set some reasonable, fact-specific limits on that very broad scope. Uniqueness as a limiter comes from the inherently duplicative nature of ESI and from FRCP 26(b)(2)(C)(i)'s admonition that discovery not be "unreasonably cumulative or duplicative." The recognition that some ESI may not need to be produced because it is "not reasonably accessible because of undue burden or cost" comes from FRCP 26(b)(2)(B)[5] (e.g., older data from legacy systems). And, the requirement that all discovery be "proportional to the needs of the case" comes from the 2015 amended[6] definition of the discovery scope itself in FRCP 26(b)(1).[7]

3.2.2 The Technological Scope of Collection

Technologically, this scope means that nothing can be overlooked based purely on its file format or its source type. If it falls within the legal scope described above, you may need to collect it to satisfy a party's request or your own information needs, regardless of whether it comes from:

- **Enterprise systems** (e.g., email, backup, or document management systems) or departmental systems (e.g., payroll, research, or compliance systems)
- **Employee computers** (e.g., organization-issued laptops or desktops)
- **Employee storage media** (e.g., thumb drives or external hard drives)
- **Employee mobile devices** (e.g., organization-issued smartphones and tablets or authorized employee-owned devices in BYOD organizations)
- **Cloud-based services** (e.g., storage services, social media services, collaboration tools)
- **Third-party service providers** (e.g., outsourced benefits management)

Collection is not necessarily limited to these common sources either. When the circumstances have warranted it, collection has been necessary from uncommon sources such as vehicle data systems,[8] wearable fitness trackers,[9] and even ephemeral data[10] (i.e., data generated and stored in memory only temporarily as part of a computer system's normal operation). Generative AI tools will certainly be the next new source, with the potential for discovery of user prompts, AI responses, and generated documents like meeting transcripts and document summaries.

As more and more devices are rendered "smart" and internet-connected, the list of potential sources will continue to grow. For example, photocopiers are almost all networked computers with internal hard drives[11] that store potentially-discoverable copies of the documents they've handled.

[4]Fed. R. Evid. 401, available at https://www.law.cornell.edu/rules/fre/rule_401.
[5]Fed. R. Civ. P. 26(b)(2)(B), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[6]Karen A. Henry and Diana Palacios, *The 2015 Amendments to the Federal Rules of Civil Procedure*, AMERICAN BAR ASSOCIATION, https://www.americanbar.org/groups/litigation/committees/minority-trial-lawyer/articles/2016/2015-amendments-to-federal-rules-of-civil-procedure/ (Mar. 1, 2016).
[7]Fed. R. Civ. P. 26(b)(1), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[8]David Horrigan, *e-Discovery Spoliation in Unusual Places: Preserve Your Pickup Truck*, THE RELATIVITY BLOG, https://www.relativity.com/blog/e-discovery-spoliation-in-unusual-places-preserve-your-pickup-truck/ (Mar. 2, 2017).
[9]Samuel Gibbs, *Court sets legal precedent with evidence from Fitbit health tracker*, THE GUARDIAN, https://www.theguardian.com/technology/2014/nov/18/court-accepts-data-fitbit-health-tracker (Nov. 18, 2014).
[10]Kenneth J. Withers, *"Ephemeral Data" and the Duty to Preserve Discoverable Electronically Stored Information*, 37 Univ. of Baltimore L. Rev. 349 (2008), available at https://scholarworks.law.ubalt.edu/ublr/vol37/iss3/4/.
[11]Federal Trade Commission, *Digital Copier Data Security: A Guide for Businesses*, https://www.ftc.gov/tips-advice/business-center/guidance/digital-copier-data-security-guide-businesses (July 2017).

3.3 HOW COMPUTERS STORE ESI

3.3.1 A Tale of Tiers and Types

To operate efficiently, computers need to be able to access and work with lots of stored information as quickly as possible:

- Some information is needed to tell a computer's components how to work together
- Some is needed to run the operating system and your applications
- Some is needed to track and respond to your inputs
- Some is needed to retain all of your activity and files

In addition, some of that information needs to be stored reliably even when the computer is off, and some of it is only needed temporarily when the computer is on and performing specific operations. Some of it never changes, and some changes all the time.

As with most things, some memory technologies are fast and expensive, while others are slow and inexpensive. Some of those technologies are volatile, requiring power to maintain storage; others are non-volatile, maintaining storage without power. A mixture of all these memory types is used to satisfy operational requirements while striking a balance between efficiency and affordability:

- **Read Only Memory (ROM)**
  - Fast, non-volatile memory that contains essential instructions for the operation of the components in the computer
- **Cache**
  - Fast, volatile memory that the central processing unit (CPU) and other computer components use to store information for rapid access to speed up tasks
  - Most computers include two levels of CPU cache, and many now include 3, as well as caches for the graphics processing unit (GPU) and the storage drives
- **Random Access Memory (RAM)**
  - Fast, volatile memory that the computer uses for temporary storage of information in active use, including parts of the operating system, parts of applications, and open user files
  - Static RAM (SRAM) is faster but more expensive and is used for the caches described above, and Dynamic RAM (DRAM) is slower but less expensive and is used for the "RAM" component of most personal computers
- **Storage Drives**
  - Slow, non-volatile memory that is used for the bulk of information storage, including the operating system, applications, and all user files and data
    - This is the memory from which collection is most often performed
  - Storage drives can be traditional hard disk drives (HDDs), which work like rewritable record players, or newer solid state drives (SSDs), which cost more but are faster and have none of the moving parts required for HDDs

- Many computers employ both types of storage drive: a faster SSD for the operating system and key software and a larger HDD for files and media
- Portions of storage drive memory may also be used as an extension of RAM, known as virtual memory, to further enhance operating efficiency

This multi-type, multi-tier approach to memory and storage is also employed in computing devices beyond laptops and desktops. For example, smartphones and tablets employ similar tiered memory systems for the same reasons.

3.3.2 Memory in Motion

As your computer or mobile device operates, there is a constant flow of information being read from and written to storage drives, RAM, and the various caches. At any given moment, multiple copies of a file or portions of a file may exist in multiple locations. These temporary copies are known as ephemeral data, since it typically only exists as long as the computer is on and the operation is active. Collections from individuals' computers and mobile devices are typically only concerned with the static ESI on the storage drive(s), but the ephemeral data generated by enterprise systems has occasionally been implicated in legal matters.[12]

3.3.3 Keeping Track of What's There

Whether a computer or mobile device is using an HDD, an SSD, or both, it is managing a collection of thousands of discrete files that is constantly evolving as files are read, modified, written, and deleted. The computer's file system dictates how this occurs, and although there are a variety of file systems in use in different types of computers and servers, the underlying principles are the same for our purposes.

The immense volume of available storage is divided up into very small physical and logical units. The smallest physical unit is typically referred to as a sector, and some common systems refer to the smallest logical unit as a cluster. The specific nomenclature and the specific relationship between physical and logical units depend on the file system in use. Regardless, the computer tracks all of those sectors and clusters in what is, essentially, an enormous spreadsheet that records what it has put where and where there is free space to put new things.

Almost all files will be large enough to occupy multiple physical sectors, but those sectors will not necessarily all be physically adjacent. Most of the time, they are spread out across the physical storage, connected only by the entries in the computer's master storage spreadsheet documenting their relationship. When files are deleted, one of two things happens, depending on the type of storage drive.

In a traditional, platter-based drive, the physical sectors are not wiped clean of their file

[12]Kenneth J. Withers, *"Ephemeral Data" and the Duty to Preserve Discoverable Electronically Stored Information*, 37 Univ. of Baltimore L. Rev. 349 (2008), available at https://scholarworks.law.ubalt.edu/ublr/vol37/iss3/4/.

fragments; rather, the master spreadsheet is just updated to delete the references to that file and to show that those sectors are available once more. In a solid-state drive, the actual data will also be deleted to prolong the life of the drive.

3.4 COLLECTING AND RECOVERING ESI FROM COMPUTER STORAGE

3.4.1 Physical and Logical Collections

As discussed above, we are generally concerned in collection with the primary, non-volatile data storage in a computer (or mobile device), whether in the form of HDDs, SSDs, or both. On SSDs, what the computer says is there and what's actually there are the same. On HDDs, there is a distinction between what's actually, physically stored on a drive and what the computer is currently tracking in its master storage spreadsheet for that drive, as noted in the last section. This results in two collection options for such drives: physical and logical.

Physical collections of HDD storage drives capture an exact copy – or image – of everything on the physical storage, regardless of what the master storage spreadsheet says about where data is and isn't on the drive. This is a bit-by-bit copy, also known as a bitstream copy, which replicates all the physical contents of the storage exactly as they are, essentially creating a virtual duplicate of that physical hardware. The primary benefits of this approach are its completeness and the potential it provides for recovery of deleted files.

Logical collections of HDD storage drives work within the file system's management of the storage drive rather than replicating the whole piece of hardware. Logical images exactly replicate everything tracked in the computer's master storage spreadsheet or some defined subset of it (e.g., everything in particular directories or folders). The primary benefit of this approach is the potential to target more narrowly and collect less extraneous material.

3.4.2 Recovery of Deleted Files

On traditional platter-based hard drives, there are two potential sources of information that can be captured in physical images that are not captured in logical ones: slack space and unallocated space. As we noted above, files in computer storage take up multiple sectors or clusters on a drive. Sectors or clusters not currently in use for active storage are referred to as unallocated space. Some sectors or clusters that are in active use may only be partially full. The remaining, unused portion of the sector or cluster is referred to as slack space.

On platter-based hard drives, computer deletion only deletes the records of what's in sectors and clusters, rather than actually erasing them, so both unallocated space and slack space may contain fragments of deleted files that had been stored there. A forensic examiner working with a full, physical image may be able to use specialized software tools to recover files or file fragments from unallocated or slack space and render them usable for investigation or litigation. While this is not a typical step in routine eDiscovery work, it can be invaluable in cases involving accidental or intentional deletion of relevant ESI.

In recent years, however, a transition has occurred from platter-based hard drives being the norm to solid-state hard drives being the norm. Most new computers and devices are now built using SSDs instead of traditional platter-based hard drives. Unlike the older drives, SSDs are designed not to retain anything in unused storage space to prolong the life of the drive. So, for any newer computers, laptops, or other devices, recovery of deleted files may not be possible.

3.4.3 Preventing Alteration During Collection

During the collection of ESI from computer or device storage drives, it is important to avoid altering the source in any way by the act of collection. As we noted above, computers are designed for efficiency rather than data preservation, and when in operation, they have a constant flow of information being read from and written to various memory components. To avoid doing any new writing to a drive during the act of reading from a drive, forensic examiners use tools called write blockers. Write blockers are specialized hardware or software tools that block any write commands from being passed to a drive while it is being accessed for collection, ensuring the original source is unaltered by the collection activity.

3.4.4 Verifying the Accuracy of Collection

Just as important as avoiding alteration to the source is verifying that the copies you've made are accurate ones. When copying large volumes of ESI (hundreds or thousands of files per source drive is common), there is some potential for errors to occur during the copying of some of those files. Hashing is used to validate that all files have been copied accurately.

Hashing is a technique by which sufficiently unique "fingerprints" can be generated for files. Hash functions are mathematical processes that take irregular-length inputs (e.g., the data in a particular file), and use them to generate fixed-length outputs (e.g., a string of 32 numbers and letters). In collection, hashing is typically accomplished using a cryptographic hash function (e.g., MD5 or SHA-1), which is well-suited to matching unique inputs to particular outputs.

To verify a collection's accuracy, one set of fingerprints is generated from the source files, and that set is then compared to a second set generated from the copied files. Fingerprint matches confirm an accurate copy, and fingerprint mismatches identify copying errors.

3.5 THE INTERSECTION OF TECHNICAL AND LEGAL REALITIES

3.5.1 Legal Realities

The ultimate goal of evidence collection is the eventual use of some of that evidence in court, whether by you or another party. The admissibility of a particular piece of evidence at trial turns on a variety of factors, including its relevance, its potential for prejudice, its status as hearsay, etc. The most foundational of the requirements offered evidence must satisfy is that it must be authentic, i.e. it must actually be whatever it purports to be. This is essential for the obvious reason that fake or falsified or altered materials cannot carry any weight as evidence; fake evidence makes no fact more or less true and is, therefore, irrelevant to the proceedings.[13]

[13]Fed. R. Evid. 401, available at https://www.law.cornell.edu/rules/fre/rule_401.

The process for establishing evidentiary authenticity is laid out in Federal Rule of Evidence 901.[14] To establish authenticity, "the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is." Satisfying this requirement for ESI means being able to demonstrate that an offered file comes from where you say it does and has not been altered from the original, *i.e.* that you've maintained forensic soundness and chain of custody.

3.5.2 Forensic Soundness

Forensic soundness is a widely used phrase in the discussion of forensic collection and investigation processes that lacks a precise legal or technical definition. It is used generally to describe tools and processes that can be relied upon to capture evidence in a way that does not alter or corrupt that evidence, and which conforms to accepted industry best practices. For working with ESI, the National Institute of Standards and Technology actually tests the operation of available forensic tools (like the write blocking and disk imaging tools mentioned above) and provides public reports on their soundness.[15]

In the context of eDiscovery, ensuring forensic soundness generally means capturing exact copies of relevant files, with any relevant metadata intact, and then working with copies of those copies, to ensure preservation of an unaltered original set. The precise technical steps required to achieve that goal will vary by ESI source and collection tools employed, and the currently-accepted industry best practices for various source types continue to evolve as the technology does, both in practice and in court. For this reason, engaging the services of a qualified forensic expert – or at least consulting with one prior to collection – is recommended to ensure currently-accepted tools and processes are employed.

3.5.3 The Importance of Maintaining Metadata

Metadata, broadly speaking, is data about data. In the context of ESI, every file on a computer or mobile device contains not only the user-facing content you would see if you opened it but also a diverse array of information about the file itself. Common examples include the time and date sent for an email, or the author and last modification date for an Office document. This additional information is the file's metadata, and it is an important part of collection and discovery in most cases.

The specific metadata fields available will vary with the specific file format. For example, music files typically include artist and track information in their metadata. Photo files may record where and by what device they were taken. Email files will document their attachments. Available fields will also vary with the source application. Application metadata ranges from the very widely-used (e.g., date and time created) to the very application-specific (e.g., tracked changes in a document or hidden content in a spreadsheet). Additional metadata about files may also come from the system on which they exist (e.g., file path).

In terms of evidentiary value, we are most often concerned with metadata revealing when things

[14] Fed. R. Evid. 901, available at https://www.law.cornell.edu/rules/fre/rule_901.
[15] *Computer Forensics Tool Testing Program (CFTT)*, NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, https://www.nist.gov/itl/ssd/software-quality-group/computer-forensics-tool-testing-program-cftt (Nov. 15, 2019).

were done (e.g., when something was created, modified, sent, or received), but there may be relevant evidence in other types of metadata, and there is enormous process value regardless. Metadata values are the basis of many filtering, sorting, and searching options within document review tools. For example, metadata is what enables you to keep emails and attachments in family groups, to filter for emails to or from a particular address, or to search for keywords within email subject lines. The more data about your data you have, the more creative and efficient you can be in your exploration of that data during early case assessment and review.

Because of both its potential evidentiary value and its enormous process utility, metadata has become an expected (and sometimes required[16]) component of many ESI productions (e.g., in DOJ productions[17]). Unfortunately, metadata is also easily altered if files are not collected and handled correctly. For example, accessing and copying original files without safeguards like those we discussed above can alter metadata, as can forwarding relevant emails[18] instead of collecting them directly, either of which would destroy forensic soundness, reduce utility, and potentially impair admissibility.

3.5.4 Chain of Custody

Chain of custody refers to documentation of the path a piece of evidence has traveled from its point of origin to its eventual submission in court. It typically documents places, times, and people involved in the handling of the evidence, as well as any relevant processes employed. Its purpose is to demonstrate that a piece of evidence submitted in court is what you claim it is, from where you claim it's from, and unaltered, as required by Federal Rule of Evidence 901.[19]

Although the concept originates with physical evidence, it is equally applicable to ESI collection and handling. Thus, carefully documenting your collection efforts and subsequent ESI handling is another important part of ensuring the reliability and later admissibility of the ESI you collect. In addition to your chain of custody documentation, an individual responsible for the collection and data handling may need to submit an affidavit (or provide live testimony) describing the steps taken, the tools used, and how forensic soundness and chain of custody were both maintained from the point of collection to the point of submission as evidence.

3.6 SELF-COLLECTION AND ITS RISKS

3.6.1 Collection Approaches: Custodian Self-Collection

Custodian self-collection refers to a collection approach in which the custodians themselves undertake the identification and collection of relevant documents from their own materials. For example, they might review their physical records and turn over any relevant paper files to a designated recipient in the in-house counsel's office, or they might review their stored electronic files and place copies of relevant materials in a designated storage area on the organization's network (or move relevant emails to a designated folder in Outlook).

Custodian self-collection of ESI carries four categories of risk that can each lead to spoliation sanctions, authentication and admissibility issues, and other negative consequences, which is what makes custodian self-collection approaches unsuitable for almost all matters:

[16] *Singh v. Hancock Natural Resources Group, Inc.*, 2016 WL 7474886 (E.D. Cal. Dec. 29, 206), available at https://scholar.google.com/scholar_case?case=2059425785764292803.
[17] Antitrust Division, *Electronic Production Letter Using Load Files*, U.S. DEPT. OF JUSTICE, https://www.justice.gov/sites/default/files/atr/legacy/2014/10/27/237704.pdf (Sept. 2014).
[18] *Singh*, supra note 16.
[19] Fed. R. Evid. 901, available at https://www.law.cornell.edu/rules/fre/rule_901.

- **Generic Inaction**: The first category of risk you run when leaving collection to the custodians is that they simply may not do it. Employees are busy doing their normal job duties, and most do not understand the importance of preservation and collection the way lawyers do. It is not uncommon to have to chase employees down just to get them to acknowledge receiving a legal hold. Asking them to execute a complex, time-consuming collection process is likely to go right to the bottom of their to-do list. And, even if you eventually get everyone to act on your instructions, the delays before action can lead to the loss or alteration of relevant materials through normal work activities, automated janitorial processes, or system or device failures.
- **Legal Misunderstanding**: The second category of risk you run when leaving collection to the custodians is that they will misunderstand or misapply the legal and factual scope information you give them in your instructions. The scope of preservation and collection is defined through the interaction of a nuanced legal standard, the pleadings and discovery requests of the parties, and the facts known at the time. The scope of relevance (and, thus, of collection) frequently evolves over the course of discovery as legal disputes are refined and more factual knowledge is gained. Expecting non-lawyer employees to clearly understand nuance with which lawyers frequently struggle is a recipe for disappointment, and expecting that nuance to be consistently applied from employee to employee is even more so. And, when employees misunderstand or misapply the scope you've tried to set, relevant materials can end up omitted or lost altogether.
- **Technical Ineffectiveness**: The third category of risk you run when leaving collection to the custodians is that – even if they perform the requested collection and apply the scope guidance as you intended – they may still execute the process in a technically ineffective manner resulting in materials being missed, lost, or altered. For example, custodians asked to run keyword searches to locate relevant materials (in email, or their local files, or enterprise systems they use) may design them ineffectively, or execute provided ones incorrectly, causing relevant materials to be missed entirely. Minor changes to search syntax or search settings can make major differences in the results returned, and syntax and settings vary from system to system. Moreover, ESI materials and their metadata are easily altered by almost any interaction with a file. Custodians working without write blockers or other forensic tools cannot maintain forensic soundness or perform hash validation. Some metadata will be altered, which may affect the ESI's evidentiary value, its authentication, or its admissibility.
- **Intentional Misconduct**: The final category of risk you run when leaving collection to the custodians is that they will engage in intentional omission, alteration, or destruction of materials to conceal their own actions. There are many situations in which your custodians' interests may run counter to your organization's. They may be responsible for some part of the events giving rise to the matter and fear getting in trouble themselves, they may be engaged in some unrelated misconduct they are afraid may be exposed, or they may think they're protecting a colleague or the organization. Whatever the reason, when custodians are trusted to self-collect ESI they have the opportunity to commit sins of omission or spoliation. And, even if they do not take that opportunity, another party may challenge the reliability of collection performed by a custodian with an individual interest in the matter or the materials.

3.6.2 Collection Approaches: Organization Self-Collection

Organization self-collection refers to an approach in which an organization leverages its information technology personnel to perform collection of ESI. For example, the administrator of the organization's email system may perform searches and exports from that system, or IT personnel might be directed to image specific employees' work computers. The materials collected by IT are then typically turned over to a law firm or a discovery services provider for subsequent processing, hosting, review, and production. While organization self-collection is a less risky approach than custodian self-collection, it is still a risky approach – and, potentially, an expensive one.

- **Legal Misunderstanding**: Just like custodian self-collection, organization self-collection carries the risk that those performing the collection will misunderstand or misapply the legal and factual scope of the matter. Achieving the proper scope of preservation and collection requires nuanced understanding of legal standards and processes, as well as the specifics of the case. As we noted above, expecting non-lawyer employees to clearly understand nuance with which lawyers frequently struggle is a recipe for disappointment and for the omission and potential loss of relevant materials.
- **Technical Ineffectiveness**: Although IT personnel will be much more technically proficient than the average custodian within an organization, proficiency with IT is not the same thing as proficiency with collection or collection tools. Thus, even if the proper legal and factual scope is applied, IT personnel may still execute the process in a technically ineffective manner resulting in materials being missed, lost, or altered, just as with custodian self-collection. IT personnel, too, may design searches ineffectively, execute provided ones incorrectly, or be unaware of specific search and export limitations inherent in different systems. Inadvertent alteration of materials or their metadata is also a very real possibility, if IT personnel are not provided with the correct tools to use or if those personnel are not properly trained in the use of those tools. Just as with custodian self-collection, such alterations may affect the ESI's usability and its admissibility.
- **Costs**: It is possible for organization self-collection to be performed effectively, but it can be expensive to do so. In addition to the costs of existing staff time or new staff altogether, many of the widely-accepted collection tools are expensive to buy or license, training and certification for the personnel who will use them add additional costs, and many of those costs must be repeated at regular intervals. Larger organizations that are frequent litigants sometimes find it cost effective to create a dedicated internal collection team, with the right tools, training, and certifications, but most organizations find this approach cost-prohibitive compared to hiring a discovery services provider as needed.
- **Delay and Disruption**: Finally, it is worth noting that periodically reassigning IT personnel from their primary duties to undertake collection activities can create delays in collection as well as disruption in the organization. Unless your organization has invested in dedicated internal collection personnel as described above, organization self-collection puts collection responsibilities in competition with existing IT responsibilities. Collection may be delayed to minimize IT disruption (creating an increased risk of loss or alteration in the meantime), or IT may be disrupted to prioritize collection (increasing the effective costs of the effort). Moreover, it is frequently necessary for those responsible for collection to later testify about the collection processes employed, which requires

additional disruption to existing IT activities (and which can be challenging for the inexperienced).

3.6.3 The Courts on Self-Collection

The risks and consequences of employing self-collection approaches are not merely hypothetical. For many years, courts have highlighted those risks, have taken parties and their lawyers to task for their reliance on self-collection in the face of those risks, and have applied significant monetary and evidentiary sanctions for failures caused by taking those risks:

- *Leidig v. BuzzFeed, Inc.*, 2017 WL 6512353 (S.D.N.Y. Dec. 19, 2017)[20]
  - In this case, an "amateurish collection of documents [led] to the destruction of perhaps critical metadata." The metadata was "irreversibly destroyed" when the plaintiff himself "transferred the files to a new device." As a result of this spoliation caused by self-collection, the plaintiff was precluded from using the dates of the affected documents as evidence.
- *National Day Laborer Organizing Network, et al., v. U.S. Immigration and Customs Enforcement Agency, et al.*, 877 F.Supp.2d 87 (S.D.N.Y. Jul. 13, 2012)[21]
  - In this case, defendant government agencies had collection searches performed by individual custodians with no meaningful direction or oversight of their searching. Moreover, most of the custodians' search efforts were undocumented, making post hoc evaluation of their adequacy impossible. As a result of this custodian self-collection process, the defendants were ultimately directed to undertake significant additional discovery work to ensure acceptable quality and completeness would be achieved. In reaching that decision, the court opined on the risks of relying on custodian self-collection approaches:

    > The second answer to defendants' question has emerged from scholarship and caselaw only in recent years: **most custodians cannot be "trusted" to run effective searches because designing legally sufficient electronic searches** in the discovery or FOIA contexts is not part of their daily responsibilities. **Searching for an answer on Google (or Westlaw or Lexis) is very different from searching for all responsive documents in the FOIA or e-discovery context**. Simple keyword searching is often not enough: "Even in the simplest case requiring a search of on-line e-mail, there is no guarantee that using keywords will always prove sufficient." There is increasingly strong evidence that "[k]eyword search[ing] is not nearly as effective at identifying relevant information as many lawyers would like to believe." [footnotes omitted; emphasis added]

- *Peter Kiewit Sons', Inc. v. Wall Street Equity Group, Inc., et al.*, 2012 WL 1852048 (D. Neb. May 18, 2012)[22]
  - In this case, the defendants were sanctioned for host of discovery failures, several of which were the result of relying on employees for management of ESI sources and for conducting the searches for relevant ESI, leading the court to conclude that the "Defendants' search of their electronic files to provide discovery responses was woefully inadequate." For example:

[20]Available at https://casetext.com/case/leidig-v-buzzfeed-inc.
[21]Available at http://pdfserver.amlaw.com/legaltechnology/National_Day_Laborer_nysd_10-3488.pdf.
[22]Available at https://www.govinfo.gov/content/pkg/USCOURTS-ned-8_10-cv-00365/pdf/USCOURTS-ned-8_10-cv-00365-14.pdf.

- "Defendants argue they had [] their most 'computer literate' employee, conduct a keyword search for 'Kiewit' from her work station and that search revealed no letters or communications with potential clients containing the term . . ."
- ". . . she is not a computer expert and has no formal training in computer sciences. Defendants have offered no evidence that they spoke with a forensic examiner or any other computer expert to even evaluate what the cost and burden of thoroughly searching Server 2."
- "[Defendant employee] claims he knows nothing about computers, but admits he dumped Server 1 in the trash, after the discovery battles had erupted and with no notice to the Plaintiff or this court, upon concluding the motherboard was 'fried.'"

Ultimately, the magistrate judge concluded that the search process employed by the defendants was not "a good faith search for the electronically stored information." As sanction, the court ordered the defendants to:

> . . . pay for Plaintiff's reasonable attorney's fees and the expenses associated with [] forensic examination of Defendants' computer equipment, and for filing and litigating the motion to compel and all other motions and hearings associated with retrieving and searching Defendants' computer files.

Moreover, due to the other evidence of defendants' "bad faith" discovery conduct, the magistrate judge recommended the application of a permissive adverse inference jury instruction.

- *SunTrust Mortgage, Inc. v. AIG United Guaranty Corp., et al.*, 2011 WL 1225989 (E.D. Va. Mar. 29, 2011)[23]
  - In this case, the plaintiff relied upon an employee central to the underlying dispute to perform identification and collection of relevant materials, and that employee took the opportunity to alter several relevant documents to make them support her version of events. In-house counsel and plaintiff executives initially discovered that the employee had materially altered two emails, but they chose not to properly investigate the employee's other ESI for further alterations. No forensics help or outside experts were hired, and after the plaintiff's IT personnel encountered some technical difficulties while trying to work with an image of the employee's hard drive, they simply gave up. As a result, in-house and outside counsel later relied upon an email they did not know had been altered to support their amended complaint. Defendants noticed a discrepancy between that version of the email and the version defendants possessed, and they enlisted an outside forensic expert who determined that the plaintiff's version had been altered, which led to significant additional discovery being undertaken.

    Ultimately, additional altered emails were discovered, and the court found that the employee had perpetrated a fraud on the court for which the plaintiff was responsible. The court also determined that "[t]he handling of the matter by [the plaintiff's] in-house counsel and its management . . . constituted an abuse of the litigation process" and that "in-house counsel, as well as its senior management, were willfully blind." As sanction for the fraud and abuse, the

[23] Available at http://amlawdaily.typepad.com/04142011suntrust.pdf.

plaintiff was ordered to pay defendant's "very significant additional legal fees and expenses" that were incurred to "preserve[] the integrity of the judicial record."

- *Green v. Blitz U.S.A., Inc.*, 2011 WL 806011 (E.D. Tex. Mar. 1, 2011),[24] vacated after settlement, 2014 WL 2591344 (E.D. Tex. June 1, 2014).
  - In this case, the defendant's discovery workflow involved a particular employee within the company meeting with counsel to find out what materials might be relevant and then talking to individuals or departments he thought might have such materials and asking them to provide him with those materials. The employee "did not institute a litigation-hold of documents, **do any electronic word searches for emails, or talk with the IT department regarding how to search for electronic documents**" [emphasis added]. As a result, numerous relevant documents were missed and never produced – some of which could have been found with "shocking . . . ease" if appropriate IT or collection experts had been involved in the process. On the basis of their significant, "willful" discovery failures, the court ordered the defendant to:
    - "pay $250,000.00 in civil contempt sanctions to the plaintiff"
    - "furnish a copy of this Memorandum Opinion Order to every Plaintiff in every lawsuit it has had proceeding against it, or is currently proceeding against it, for the past two years" (or pay an additional $500,000 sanction if they failed to do so)
    - "for the next five years . . . in every new lawsuit it participates in as a party, whether plaintiff, defendant, or in another official capacity, it must file a copy of this Memorandum Opinion and Order with its first pleading or filing in that particular court"

3.7 IN-PERSON AND REMOTE COLLECTIONS

3.7.1 Collection Approaches: In-Person Collection

Traditionally, the most common collection approach has been in-person collection, in which the person executing the collection is in physical possession of the device(s) to be collected. This may be achieved by sending the devices to the person executing the collection, but is more often achieved by having the person executing the collection travel to where the custodian(s) and their device(s) are.

In-person collection has many benefits. Most importantly, it ensures proper collection from the original source overseen by a professional rather than by the custodian. It can also give the person executing the collection an opportunity to interact with the custodians to gather useful information about the sources and what they contain, combining in-person custodian interviews with collection itself. And, when multiple custodians are in the same office location, it can be efficient, even when travel to that location is required.

In-person collection is not ideal for all situations, however. When custodians are distributed across multiple office locations – or when many employees work remotely from home –

[24] Available at https://casetext.com/case/green-v-blitz-usa.

travel to all of those locations can quickly become too costly and time-consuming to make sense. Having one or more collection professionals on-site can also be disruptive to normal operations and may not be as subtle an approach as you require in certain investigative contexts.

3.7.2 Collection Approaches: Remote Collection

As geographically-distributed (and remote) employees have become more common, remote collection has grown in popularity as an approach. Remote collection comes in four primary subtypes: self-executing devices with instructions, preconfigured drives plus remote access, preconfigured laptops plus remote access, and enterprise applications:

- **Self-Executing Devices with Instructions** – In the first subtype, a collection device is prepared in advance and shipped to the custodian who follows provided instructions to connect the device to their computer and initiate the pre-defined collection process. When the collection is complete, the custodian returns the device.
- **Preconfigured Drives Plus Remote Access** – In the second subtype, which is the most widely used, a preconfigured drive is shipped to the custodian who connects it to the computer and then grants remote access to a remote collection professional to execute and oversee the actual collection to the drive. When the collection is complete, the custodian returns the drive. (It is also possible for limited collection over the internet to be done by remote access, but typically the sizes involved make shipping drives a better choice for this approach.)
- **Preconfigured Laptops Plus Remote Access** – The third subtype is a solution developed for remote collection of smartphone data. In this approach a laptop with the necessary collection software is shipped to the custodian, who connects the laptop to the internet and their smartphone to the laptop. A collection professional can then connect to the preconfigured laptop remotely to perform the required collection. When the collection is complete, the custodian returns the laptop.

- **Enterprise Applications** – In the fourth subtype, an enterprise application is installed on the organization's network environment that facilitates manual or automated collections from devices connected to the network. Collections may be executed with or without the custodians' knowledge and may be administered by IT personnel or by third-party collection professionals. Collections are copied over

the network and later may be transferred to drives for shipping to a third-party eDiscovery services provider for processing and review. Due to the cost of such applications, they are most often used by large organizations.

3.8 OTHER IMPORTANT COLLECTION SOURCES

Thus far, we have spoken primarily about the collection of ESI materials from the computers of individual custodians, but most cases involve collection from a range of other sources as well. The fundamentals of computer memory operation and successful acquisition from that memory are the same regardless, however you still need to be aware of the other source types you may need to consider and the complications that they entail.

The other major categories of sources are:

- Enterprise systems
- Mobile devices
- Social media sources
- Cloud sources

3.8.1 Collection from Enterprise Systems

Enterprise systems refers to the software and hardware systems maintained by your organization or its departments, including email systems, internal instant messaging systems, document management systems, CRM or ERP systems, internal collaboration tools, backup systems, and more. Depending on the nature of the matter, it might also include voicemail systems, security and video systems, or even networked photocopiers or other office machines.

How collection from such systems is performed can vary widely depending on the system. Some systems store their data in ways that can be directly collected like the materials on a custodian's computer, while others require you to use the system's built-in search and export tools. Those tools may have material limitations that affect what results a search can return or what an export can contain. Working closely with the responsible IT personnel to ensure those limitations are understood and accounted for is critical when collecting data from enterprise systems.

3.8.2 Collection from Mobile Devices

Mobile devices – smartphones in particular – have become ubiquitous for both personal and business life. Like all consumer technology, there are a plethora of models and types available, and new ones are released by each maker each year. And, because many organizations have adopted bring-your-own-device policies (BYOD), organizations may have a much wider variety of smartphones as potential sources than computers (which still tend to be organization-selected and issued).

Smartphones are more difficult, more costly, and more time-consuming to collect and process than computers. The difficulty, cost, and time can vary from model to model, from maker to maker, and from operating system to operating system. Collection directly from smartphones requires specialized tools like those used to collect from a custodian's computer. Collections instead from cloud-based backups of the smartphone in question are sometimes also an option.

Additionally, it is important not to overlook less common mobile devices that may, at times, be relevant, such as vehicle GPS or data systems, wearable devices like fitness trackers, etc.

3.8.3 Collection from Social Media Sources

For better or worse, social media is currently an influential, indispensable part of American life. As it permeates its way ever deeper into our professional and personal lives, its impact upon discovery is growing in parallel. There are three main options for the acquisition of social media materials for use in litigation:

- *Printing out the material or capturing a screen image of it* – this is fast and inexpensive but does not capture any native files or metadata, and it may create authentication and admission problems down the road
- *Using the self-service export tools provided by the social media platform* – this, too, is fast and inexpensive, but it also may not provide native files or metadata, and it may come as a single, massive PDF that must be parsed out into individual records
- *Using specialized forensic collection software like that you would use to collect from a custodian's computer* – this carries additional costs (either for such a tool or for services from a provider with such a tool), but it can be essential for cases involving large quantities of social media materials, questions best resolved through the materials' metadata, or the potential for disputes over the authenticity and admissibility of the social media materials themselves

3.8.4 Collection from Cloud Sources

The final category of sources you may encounter today is cloud-based sources. These are cloud-based services used either by the organization or by the individuals within it. Examples include email solutions like Gmail, storage and sharing solutions like Dropbox, collaboration tools like Slack and Teams, and office suites like Google Docs and Microsoft 365. Much as with enterprise systems, collection from cloud sources is very dependent on the specific source and the features it includes. Some common cloud sources have extensive tools geared towards discovery activities (e.g. Microsoft 365), while others have much more limited search and export options. Successful collection from cloud sources typically requires the assistance of an experienced collection expert (as well as the cooperation of the account holder, for individual's accounts), and it may require custom solutions.

In recent years, collaboration tools have become one of the most commonly-needed sources for discovery. From its launch in 2013, Slack was the fastest-growing workplace software ever, topping 500,000 daily users in 2015.[25] Despite Slack's four-year head start, by November 2019, Microsoft Teams had surpassed Slack with 20 million daily users.[26] This rapid growth was turbocharged by the pandemic and consequent shift to remote and hybrid work, which resulted in geometric growth for Microsoft Teams. Daily active users tripled in 2020, and then they more than doubled again in 2021. By the end of 2022, Teams had over 270 million monthly active users.[27]

[25]Jack Linshi, "This 1-Year-Old Startup Says It's the Fastest-Growing Business App Ever," Time (Feb. 12, 2015), available at https://time.com/3705218/slack-business-app/.
[26]Mary Jo Foley, "Microsoft says it has 20 million daily active Teams users," ZDNET (Nov. 19, 2019), available at https://www.zdnet.com/article/microsoft-says-it-has-20-million-daily-active-teams-users/.
[27]Lionel Sujay Vailshery, "Microsoft Teams: number of daily active users 2019-2022," Statista (Jan 13, 2023), available at https://www.statista.com/statistics/1033742/worldwide-microsoft-teams-daily-and-monthly-users/.

3.9 KEY TAKEAWAYS

There are eight key takeaways from this chapter to remember:

1. Understanding the fundamentals of collection is necessary to successfully navigate discovery and to fulfill lawyers' duty of technology competence.
2. The potential scope of collection is very broad, both legally and technically, and it continually evolves as new devices and services become available and as people's patterns of behavior change to incorporate those things into their work and their lives.
3. The complex nature of computer operations and ESI storage leads to the creation of unneeded duplicates and ephemeral files, as well as the potential for easy alteration and accidental loss (though deleted ESI may sometimes be recoverable).
4. Ensuring that forensic soundness (*i.e.*, no alternation) and chain of custody (i.e., documented path from original source to introduction in court) are maintained during collection is essential to avoiding potential issues with authentication and admissibility
5. Metadata has both evidentiary and process value, and it is an expected – and sometimes required – component of collection (and later production).
6. Collecting ESI, without alteration and without loss of metadata, requires special tools (e.g., write blockers) and processes and, typically, outside experts in forensic collection.
7. Available collection approaches include: custodian and organization self-collection (high risk), in-person collection (sometimes high cost), and several varieties of remote collection (currently the most popular).
8. When planning collection, don't forget source types beyond individual custodians' computers, including: enterprise and departmental sources, mobile device sources, social media sources, cloud sources, and more (e.g., vehicle systems, wearables, etc.).

Chapter 4

Time to Make the Donuts: Processing Fundamentals

About this Chapter

In this chapter, we will review the fundamentals that legal practitioners need to know about ESI processing to successfully navigate this phase of discovery and to effectively work with their internal or external service providers to do so, including: essential steps and tools, common errors and special cases, objective culling options, and final steps.

4.1 TIME TO MAKE THE DONUTS

The range of potential electronically-stored information (ESI) sources is continually multiplying and diversifying. In order to work with that diverse range of materials during assessment, during review, at depositions, and at trial, we need a way to avoid using as many different pieces of software as there are types of sources and a way to enable searching and document identification across different source types. It is for these reasons that we need processing. Moreover, ESI processing for discovery is one of the areas in which legal practitioners need some level of understanding to fulfill their duty of technology competence for eDiscovery.

Although it is often given short shrift compared to the steps that come before it (identification, preservation, and collection) and after it (assessment, review, and production), effective processing is critical to the success of those downstream steps and includes a variety of important technical decisions that can have substantive effects. For example, if processing is not performed correctly, or if the wrong decisions are made, searches can be rendered unreliable, materials can be rendered unusable, and production options may be affected.

4.1.1 Processing Specifications

The importance and complexity of processing can be seen in the level of detail involved in written production specifications (which can effectively dictate how processing needs to be performed to facilitate the required production). Although not all cases involve a written set of production specifications, many do – some developed through negotiation and others dictated by agencies or courts. These specifications may cover a wide range of process details, such as:

- The time zone in which dates and times should be presented
- The metadata types that must to be captured
- The application of optical character recognition
- The visual parameters for generating document images
- The handling procedures for special data types

Any one of these details could have impact the usability of the processed materials and the effectiveness and efficiency of downstream discovery activities – both for your work with your own materials, and for your work with the materials that are produced to you by other parties.

4.2 KEY ACTIVITIES AND COMMON TOOLS

Broadly speaking, there are four main activities that take place during processing: expansion, extraction and normalization, indexing, and objective culling. To start, we will discuss the first three of these activities. The fourth activity, objective culling, will be covered separately below.

4.2.1 Expansion

The first thing that must be accomplished when processing ESI for discovery is the expansion of all containers. Container files are files that contain other files within them. Common end-user examples would include PST files, which contain countless individual email files, or ZIP

files which can contain any combination of file types. Additionally, the collection process often generates container files that must also be expanded (e.g., hard drive images). Beyond just container files that contain groups of files, many file types can also contain individual embedded objects.

Examples of embedded objects would include an attachment to an email, a logo image in an email signature, or a spreadsheet chart embedded in a Word document. Embedded objects can be handled in different ways, and it is common to separate some to handle as discrete files and to leave others embedded. For example, it is common to extract all email attachments as their own files (though tracked in family groups with their parent email), and it is common to leave in-line images in documents embedded to be reviewed in context. Decisions about how to handle different kinds of embedded objects can affect both searchability and reviewability.

4.2.2 Extraction and Normalization

As we noted above, ESI for discovery can come in hundreds of different file formats that could each require a different piece of software to natively view. To avoid litigants on either side from having to utilize all those different source applications, and to enable integrated management, searching, and review across file types, ESI must undergo a process of extraction and normalization. In this process, all of the content is extracted from your collection of electronic files, and that extracted content is normalized into a consistent, usable format.

The content to be extracted includes all of the text from the files (e.g., the body of an email or a Word document), and may include any objects embedded within the documents. For files with imaged text (e.g., scanned documents), it is common for optical character recognition (OCR) to be used to attempt extraction of the available text. Beyond a file's primary textual content, its metadata must also be extracted (e.g., created by, date modified, etc.), and its relationships to any other files must be documented (i.e., if it is, or has, an attachment).

All of this extracted content is normalized into database fields that can be displayed consistently by document review software to facilitate downstream discovery activities like early case assessment (ECA) and review. Decisions may need to be made at this stage about how hidden content in documents should be handled and what version of a document and its text should be extracted and displayed (e.g., tracked changes in Word, speaker notes in PowerPoint). Typically, the native files are still available too, linked to their extracted, normalized content, so that the native versions can be viewed instead when necessary. For example, the extracted text from a spreadsheet is not very comprehensible compared to the native version of that spreadsheet.

4.2.3 Indexing

In order to make all of those downstream activities possible, all of the extracted content also needs to be indexed. Indexing is the process of creating the enormous tables of information that are used to power search features. Most common are inverted indices, which essentially make it possible to look up documents by the words within them. Inverted indices are like more elaborate versions of the

indices you find in the backs of books. Decisions about how indices should be generated and what common words (e.g., articles, prepositions) they should skip affect the completeness of search results later. Searches can only find what indexes show.

More sophisticated semantic indices are created to power features like concept searching, concept clustering, and technology-assisted review. These multi-dimensional indices come in a few varieties, but essentially, they all document how frequently words appear in the same documents as other words. From these co-occurrences, the tools can identify clusters of related terms that define a topical area and transcend any one, specific keyword. Some customization of this process is also possible, which will affect what results those indices reveal.

4.2.4 Common Tools

All of these core processing activities are completed using specialized software tools. These tools are able to automatically perform expansion, extraction, normalization, and traditional indexing for many common file types, and they can generally be operated manually or customized as needed to handle more challenging or less common file types. How wide a range of common file types and issues can be handled automatically varies from tool to tool, as does how easily (and to what extent) custom solutions can be implemented to go beyond that range.

Those tools might come in the form of off-the-shelf processing software from third-party providers. They might come integrated with a document review platform or with an enterprise collection platform. Because of the frequent need for adaptability, customizability, and scalability in processing activities, many eDiscovery service providers develop proprietary processing tools to handle these activities. Such organizations may still use accepted third-party tools like to perform the necessary indexing for search.

4.3 COMMON EXCEPTIONS AND SPECIAL CASES

Although a great many of the files encountered during ESI processing are common types that can be handled in a standardized, automated way, not all are. Almost every processing effort encounters at least a few exceptions during processing that cannot be handled without some manual intervention (if they can be handled at all). Additionally, certain source types are special cases that routinely require custom work to process. The handling of these exceptions and special cases can affect both project costs and the completeness of your data set.

4.3.1 Common Processing Exceptions

During the processing of a data set, the processing tool being used will often encounter some files that it cannot automatically process. Most commonly, these exceptions occur because the tool has encountered either corrupt data, a password protected file, or an unknown file type. In any of these cases, the exception is logged and reported for later review by the processor and, when needed, by the case team.

Corrupt Data

Corrupt data is data that is physically incomplete or unreadable in some way. This can be the result of a faulty sector in the storage medium, the result of a copying error at some point in the chain of custody, or the result of the original source itself having been corrupted. In some cases,

content can be recovered from corrupted file data, but in many cases, retrieval of a non-corrupt version from the original source is required – assuming that source was not corrupted.

Password Protected Files

Many file types – from PST files to PDF files to Microsoft Office documents – can be protected with passwords by their authors or owners. Files that are protected by passwords cannot be automatically processed until they are unlocked so that their content can be extracted. In some cases such passwords can be cracked or bypassed through the use of specialized tools, but in most cases getting the password from the author or owner is the fastest, cheapest solution.

Unknown File Types

Although most processing tools can identify and handle hundreds of common file types, there are thousands of file types in existence – including countless proprietary file types generated by custom corporate tools and systems. When a processing tool encounters a file of either a type it doesn't recognize, or for which it cannot identify a type at all, it logs an exception for later review by the processor. If the type is unknown, manual expert review may be able to identify it, if the type is known but uncommon or proprietary, custom work may be required to process it – if it can be processed at all.

4.3.2 Source Type Special Cases

In addition to the unpredictable occurrence of processing exceptions, there are a range of source types that will almost always require some custom work and some additional decisions to handle. Examples include:

- **Mobile Devices** – data from mobile devices frequently contains databases and other aggregated types of data that require custom unitization into smaller chunks or discrete records, for example: contacts databases, message threads, email databases, call logs, etc.
- **Social Media** – data from social media websites and applications can, like mobile device data, include aggregated data types that must be unitized into smaller chunks or individual records, plus other challenges such as: linked content, in-line graphical elements like emoji, and specialized metadata fields documenting reactions and other details
- **Collaboration Tools** – collaboration tools such as Slack and Teams potentially contain enormous volumes of individual messages, organized into a wide variety of threads and channels (requiring unitization), and integrate a significant volume of embedded and linked content (causing collaboration tool data to expand far more in size during processing than most types), plus in-line graphical elements and specialized interaction fields like social media
- **Backup Tapes** – working with materials from backup tapes can require restoration of large volumes of data to obtain the desired files for processing; with some tools, tapes can be indexed prior to restoration to allow for targeted restoration instead
- **Structured Data** – structured data includes the large operational databases that underpin corporate systems like CMS or ERM, and custom work is generally required to identify and capture the right portion of that data and then present it in a usable form

4.3.3 Unitization

As noted above, mobile device sources, social media sources, and collaboration tool sources all raise questions of unitization. These source types frequently include ongoing threads of back-and-forth messages (e.g., text message threads, direct message threads, Slack channel threads, etc.), which can span long periods of time. Although the specifics vary by source, these message threads are often maintained in ongoing logs that are not conducive to efficient review or later use as evidence. Rather than present weeks or months of messages in a single document, it is typical to unitize these logs into separate, shorter documents for review a nd production.

When doing so, some judgment must be exercised about what size the units should be. Individual messages stripped of thread context are also not ideal (as courts have pointed out[1]), so some middle ground between massive logs and single messages is preferred. It is common to unitize such materials into 24-hour chunks, so that each day's communications become a single document, but other divisions may be rational depending on your materials and case.

4.4 OBJECTIVE CULLING OPTIONS

In addition to the above activities, processing also includes several types of objective culling that are used to reduce the amount of material that must be worked with throughout the subsequent phases of a discovery project, saving both time and money. The objective culling options commonly employed during processing are de-NISTing, deduplication, and content filtering. It is important to understand the operation and limitations of these culling options so that you can make informed decisions about how to deploy them in your projects.

4.4.1 De-NISTing

Basic de-NISTing is a standard step performed in almost all discovery processing efforts. De-NISTing removes the known files that make an operating system or a software program work, such as executables, device drivers, initialization files, and others, which together can make up more than half the volume captured in a drive image. The trend towards more targeted collection methods and fewer full images has reduced the impact of de-NISTing on overall volume somewhat, but it is still an important step to remove material certain to be irrelevant.

De-NISTing does not do this by file types or file names, but by identifying specific, known files that match those originally released by the developer – thereby ensuring that they are free of user alteration or content and are, therefore, irrelevant. The process works using hashing to compare collected files to lists of known system files.

Hashing is a technique by which sufficiently unique "fingerprints" can be generated for files. Hash functions are mathematical processes that take irregular-length inputs (e.g., the data in a particular file), and use them to generate fixed-length outputs (e.g., a string of 32 numbers and letters). In collection, hashing is typically accomplished using a cryptographic hash function (e.g., MD5 or SHA-1), which is well-suited to matching unique inputs to particular outputs. Identical files produce identical hash values, and hash values can be easily compared by software to automatically identify matches.

[1] *See, e.g., Laub v. Horbaczewski,* 331 F.R.D. 516 (C.D. Cal. Apr. 22, 2019) (Magistrate Judge expressing a preference for "aggregated" formats preserving "the integrity of the threads of communication reflected in the text messages"), available at https://casetext.com/case/laub-v-horbaczewski.

The "NIST" in the name of this culling option is an acronym for the National Institute of Standards and Technology, which runs the National Software Reference Library[2] (NSRL), which is the source of the Reference Data Sets used for this process. Their lists of hash values for known file are updated a few times per year, but because of the volume and diversity of software in use today, and the frequency with which software is updated, those lists are rarely complete enough to remove all of the irrelevant software and system files that are present.

To compensate, some supplemental filtering is often performed along with de-NISTing. This supplemental filtering may be accomplished using either "stop filters" (also called "exclusion filters") or "go filters" (also called "inclusion filters"). Stop filters exclude specified file types and leave everything else included, while go filters do the opposite, excluding everything and leaving only specified file types included. The difference is what happens to your unknown unknowns (i.e., stop filters let them pass through to later steps and go filters eliminate them at this point).

4.4.2 Deduplication

Deduplication leverages the same hashing process described above for de-NISTing, except used now to compare the collection against itself rather than against an outside file list. The operation of computers and enterprise information systems naturally produces many identical copies of files and messages in many different places that add no additional relevant information. Hashing a collection and scanning the values for those duplicates typically allows for a significant volume reduction and is a standard discovery processing step for this reason.

Deduplication can be performed on either the file level or the family group level. The family group level is far more common because the integrity of family groups is typically maintained throughout discovery. For example, if the same spreadsheet was attached to two different emails, neither copy would be removed, but if two copies of the same email with the same spreadsheet attachment were present, one of those identical family groups would be removed.

Deduplication can also be performed on either the custodian level or globally. Global deduplication is far more common (since it is possible to programatically track everywhere a duplicate was for later reporting or restoration as needed.)

4.4.3 Content Filtering

Finally, processing typically affords you the option to perform some content filtering, including date range filtering and keyword filtering (and potentially filtering based on certain metadata fields and values). Date range filtering at this point in a discovery project is very common, as there are often clear temporal limits on relevance that can be applied. There are still some specifics of which to be aware, however:

- Files typically carry multiple date/time stamps that can vary by file type, potentially including date created, date last modified, date last accessed, date sent, and others, giving you choices of which to use. For this reason, it is common for service providers to create a custom master date metadata field that draws from different date/time fields for different source file types (e.g., pulling date/time sent

[2] *National Software Reference Library (NSRL),* NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, http://www.nsrl.nist.gov/index.html (Nov. 18, 2019).

for emails and date/time last modified for documents so that they can be sorted and filtered together).

- If you are maintaining family groups, as is typical, you will need to consider whether parent dates override child dates for filtering, or whether any file in a family group can pull in the whole family group (e.g., does an email from outside your date range with an attached file that's from inside your date range get included?).
- Because of computer errors or collection issues, some files may also have inaccurate, impossible, or nonexistent date stamps, so you may need to manually check out any outliers in your chosen date field (e.g., 1/1/1900 or 0/0/0000) to see what they are.

Keyword filtering is done less often at this point in a project, because the available search and analysis tools are often less robust than those available during later phases and are generally not directly accessible by the case team (requiring a technician to execute searches and report back, making iterative testing and improvement cumbersome). However, in cases with keywords that are fixed by negotiation or court order, it can be safe and useful to employ.

4.5 FINAL STEPS

In addition to the four core activities of expansion, extraction and normalization, indexing, and objective culling that we have already discussed, there can be a variety of additional steps required during processing to prepare the materials for subsequent early case assessment, review, and production activities.

4.5.1 Potential Additional Steps

Depending on the platform in which the material will be used and the ways that it will be used, additional steps may be required to finish preparing it for those activities. For example, we noted above that it is common to create and populate a custom master date field that integrates values from different date/time fields associated with different file types. It is also common to create other custom metadata fields, such as a field that extracts the domain names associated with email addresses, or a field that documents collection source details such as custodian or directory. The specific fields to be created will depend on the material with which you will be working and what you hope to accomplish with it during ECA and review.

In addition to custom metadata fields, final preparation activities may also include the preemptive generation of TIFF images of the documents (i.e., PDF-style page images), if there is a desire to review documents in that form (or a need to have them ready for rapid production turnaround later). And, if the subsequent activities are taking place in a different software platform than the processing (which is often the case), some form of load file will also need to be prepared to facilitate the review platform's ingestion of the processing platform's output.

Load files are, essentially, enormous tracking spreadsheets that can contain every document, its extracted metadata (and any custom fields), its extracted text content, links to associated native files, links to standalone text files, links to associated TIFF images, and other details. They serve as Rosetta Stones for the ECA and/or review software to understand how all the thousands upon thousands of discrete files and pieces of information you're loading into it for a given project fit together in a usable way.

4.5.2 Validation

Regardless of the specific steps taken in a given processing project, all processing efforts generally end with some form of quality control validation process prior to the hand-off to ECA and review activities. Given the enormous volume, diversity, and complexity of materials resulting from processing, a wide range of simple technical issues are possible, including file naming errors, load file field errors, file linking errors, imaging errors, and more. To identify such issues prior to loading for subsequent activities, processors typically employ some combination of targeted quality control checks for specific issues, random sampling checks to spot any other issues, and software validation tools to backstop the human checks.

4.6 KEY TAKEAWAYS

There are five key takeaways from this chapter to remember:

1. Having at least a basic understanding of processing activities is essential to fulfilling a lawyers' duty of technology competence for eDiscovery, because processing decisions can and do have substantive effects on downstream discovery activities.
2. The primary processing activities are: expansion, which opens container files and handles embedded objects; extraction, which captures text and metadata from all those files; normalization, which standardizes the format and appearance of that extracted content; and indexing, which creates the inverted and semantic indices that enable searching.
3. It is common to encounter materials during processing that require custom work or that cannot be processed at all because they are password protected, corrupted, or of an unknown type; moreover, some source types will almost always require custom work, like mobile devices, collaboration tools, and social media.
4. During processing, objective culling is typically performed to remove system files (de-NISTing and file-type filtering) and to remove duplicates (deduplication); additionally, you usually have the option to perform content filtering based on date range (very common, especially if negotiated) and keyword (less common, due to process limitations).
5. Additional steps will need to be taken to finish preparing your processed materials for ECA and review, potentially including custom field creation, TIFF image creation, or load file creation – depending on your goals and tools, and always including some form of quality control validation process to check for common technical errors.

Chapter 5

Clearing the Fog of War: ECA Fundamentals

About this Chapter

In this chapter, we will review the fundamentals that legal practitioners need to know about performing effective ECA in the context of eDiscovery, including: sampling, searching and filtering; threading and duplicates; advanced analytic tools; integrated workflows; and more, to equip you “to scent out the truth.”

5.1 CLEARING THE FOG OF WAR

In the early nineteenth century, the Prussian military analyst Carl von Clausewitz wrote of the overwhelming uncertainty[1] inherent in decision-making during military conflicts:

> War is the realm of uncertainty; three quarters of the factors on which action in war is based are **wrapped in a fog of greater or lesser uncertainty**. A sensitive and discriminating judgment is called for; a skilled intelligence to **scent out the truth**. [emphasis added]

By the end of the nineteenth century, the phrase "fog of war" had appeared as a shorthand description for this state of uncertainty, and this apt metaphor has remained in wide use to the present day, applied in every context from business to video games.

5.1.1 The Fog of Litigation

The fog of war is also apt shorthand for the state of uncertainty that exists early in a new legal matter. Whether you are gearing up for litigation, an agency enforcement action, or an investigation, you are faced with potential conflict and liability shrouded in a fog of uncertainty:

- What are the actual underlying facts?
- What are the legal and financial risks?
- What evidence might exist? With us? With them?
- How strong or weak of a position are we in overall?
- What will it cost to proceed, and how much should we spend?

Early case assessment (ECA), fundamentally, is the process of trying to clear some of the fog of uncertainty around the answers to these questions and others like them. As litigation has evolved in the eDiscovery era, however, so too has the scope of what's included in ECA.

5.1.2 Three Goals, One Name

The name ECA is now used to encompass not only the traditional ECA described above, but also what might better be called Early Data Assessment (EDA) and Downstream Activity Preparation (Downstream Prep). All three goals are attempts to remove some of the fog by reducing one area of uncertainty:

- **Traditional Early Case Assessment**
  - Traditional ECA was and is focused on reducing uncertainty about the risks and costs associated with a new legal matter to inform a decision about how to proceed; because doing so now requires reviewing relevant ESI, eDiscovery processes have become essential for effective traditional ECA.
- **Early Data Assessment**
  - EDA is a new goal that has been lumped under the ECA banner and is focused on evaluating the contents, properties, and completeness of collected ESI materials to reduce uncertainty about ongoing preservation, collection, and processing decisions, among others.

[1] *Fog of war,* WIKIPEDIA, https://en.wikipedia.org/wiki/Fog_of_war (last visited July 22, 2021).

- **Downstream Activity Preparation**
  - Finally, Downstream Prep has also been lumped under the ECA banner and is focused on reducing uncertainty about downstream attorney review and eventual production, including testing and refining searches and filters, evaluating potential workflows, and estimating needed resources.

The intersection of these three connected-but-distinct goals can make the ECA phase of an eDiscovery effort a confusing one for practitioners. What should I and others be doing to accomplish these goals, and how should we be doing it? How do we start to clear the fog?

5.2 SAMPLING TOOLS AND TECHNIQUES

In almost any modern document review platform, case teams have a powerful set of tools at their disposal for investigating their collected ESI. The specific bells and whistles of those features vary, but the core functions almost always include: searching tools, email threading tools, duplicate handling tools, advanced analytic tools, and random sampling tools. Effective ECA involves leveraging as many of these features as are helpful – like aligning a series of overlapping lenses, to bring your quarry into sharp focus. Which lenses are helpful will depend on the specifics of your matter, your ESI, and your available time and resources.

We will discuss each of these core functions in turn and how they can be leveraged for the three goals of ECA, beginning with sampling.

5.2.1 Sampling

One of the most powerful tools in your ECA toolkit is sampling. There are a lot of ways to find materials you expect to be in a collection of ESI, but sampling is a terrific way to also find materials you didn't know to look for: the unknown unknowns. For our purposes, sampling comes in two flavors: judgmental sampling and formal sampling.

Judgmental sampling is the informal process of looking at some randomly selected materials to get an anecdotal sense of what they contain, whether that's sampling from a particular source, from a particular search's results, or from a particular time period. You're not reviewing a particular number of documents or taking a defined measurement with a particular strength; you're getting an impression and making an intuitive assessment.

Formal sampling is just the opposite: you are reviewing a specified number of randomly-selected documents with the goal of taking a defined measurement with a particular strength. Typically, that measurement is either of how much of a particular thing there is within a collection (i.e., estimating prevalence) or of how effective a particular search is (i.e., testing classifiers).

- **Estimating Prevalence**
  - Estimating prevalence is the process of reviewing a simple random sample of a given collection of materials to estimate how much of a given kind of thing is present. You might estimate the prevalence of relevant materials, of privileged materials, or of materials requiring redaction or other special steps. The size of the sample you need is dictated primarily by how precise you want your estimate to be (i.e., margin of error), and how certain about it you want to be (i.e., confidence level), and to a lesser extent, by how large your collection of materials is (i.e., sampling frame). Most often you will be dealing with sample sizes of a few thousand (e.g., a sample of 2,345 for a confidence level of 95% and a margin of error of +/-2% in a collection of 100,000 documents).
- **Testing Classifiers**
  - Testing classifiers is the process of seeing how effective and efficient a particular classifier – typically a search of some kind – actually is. Using this technique, you can estimate how much of what you're seeking a given search is likely to return (i.e., recall) and how much irrelevant material is likely to get returned with it (i.e., precision). These measurements are taken by running the searches against a control set, which is made be pre-reviewing and coding a sufficiently-large random sample. Comparing the search results to the already-completed coding allows for the iterative refinement of searches to increase their recall and precision before they are applied to the full collection.

5.2.2 Sampling and the Three Goals

Sampling supports traditional ECA by helping you rapidly learn about the materials you have, what might be in them, and how best to surface more of what you need – all without the risk of missing the unknown unknowns that can come with relying on searching alone. Reviewing a randomly-selected cross section lets you see some of everything you have and some of all the different terms and phrasing your custodians use to help you better plan your next ECA steps.

Beyond just reviewing the random sample for traditional ECA, leveraging prevalence estimation is invaluable for EDA and for Downstream Prep. Accurately estimating what you have enables you to: (a) prioritize the materials you have; (b) find gaps requiring further collection; and, (c) estimate your needed project resources, optimal project workflows, and likely project costs and durations (including assessing the viability of a technology-assisted review (TAR) or continuous active learning (CAL) solution, or the need for additional objective culling). As projects progress, prevalence estimations can also provide a yardstick against which to measure progress and completeness.

Testing classifiers, too, benefits EDA and Downstream Prep. Imagine iteratively refining searches for your own use, or negotiating with another party about which searches should be used, armed with precise, reliable information about their relative efficacy. Courts consistently prefer arguments and negotiations based on actual information to those based merely on conjecture and assumption and have emphasized the value of sampling in many[2] contexts.[3] And, the more effective the searches you develop, the more you reduce the volume of material left for downstream review and production, along with the costs of those activities.

[2] *See, e.g., Victor Stanley Inc. v. Creative Pipe Inc.*, 250 F.R.D. 251 (D. Md. 2008), available at https://casetext.com/case/victor-stanley-inc-v-creative-pipe.
[3] *See, e.g., City of Rockford v. Mallinckrodt ARD Inc.*, 326 F.R.D. 489 (N.D. Ill. 2018), available at https://casetext.com/case/city-of-rockford-v-mallinckrodt-ard-inc-1.

5.3 SEARCH AND FILTERING TOOLS

After sampling, the next major category of tools and techniques available is search and filtering, including keyword and phrase searching, Boolean searching, fuzzy searching, conceptual searching, and more.

5.3.1 Searching

Searching, both on the internet and among our own emails, messages, and files, has become an inescapable part of everyday life. Almost all of this searching, like the searching you do in eDiscovery, is powered by some form of indexing. In the eDiscovery context, indexing is typically performed during the processing phase of the project.

Indexing is the process of creating the enormous tables of information that are used to power search features. Most common are inverted indices, which essentially make it possible to look up documents by the words within them. Inverted indices are like more elaborate versions of the indices you find in the backs of books. Decisions during processing about how indices should be generated and what common words (e.g., articles, prepositions) they should skip affect the completeness of search results you get during ECA. Searches can only find what indices show.

More sophisticated indices are created to power features like concept searching, concept clustering, and technology-assisted review, which we will discuss further below in our section on advanced analytic tools and techniques.

The types of indices that are prepared and the specific features your software offers for working with them will dictate what types of searching are available to you.

- **Keyword and Phrase Searching**
  - Exactly as it says on the tin, keyword and phrase searching lets you search for a key word, for a phrase, or for lists of both at once. Just as with the basic internet searching we all use, if one of the desired keywords or phrases is present, the document will be returned. One key area of variation from tool to tool is whether wildcard characters can be used to find variations on words and, if so, how they can be used.
- **Boolean Searching**
  - Boolean search is the next step up in sophistication from basic keyword searching. It allows the use of operators such as “and,” “or,” and “not.” These operators allow for the searcher to define specific relationships between key words and phrases to achieve higher quality results (i.e., improved recall and precision). Other operators may be available, including proximity operators (i.e., to find a particular word appearing within a certain number of words of another particular word).

    The range of specific operators available varies with the tools being used, as can their precise operation. Thus, it is important to understand the tools you are actually using to be sure you are searching the way you intend.

- **Fuzzy Searching**
  - Fuzzy searching (also sometimes referred to as approximate string matching or stemming) is another extension of basic keyword searching that may be available to you. Fuzzy searching allows a search to return variations on a word rather than just the precise word you searched (e.g., finding both invite and invitation). How much variation is allowed is typically an adjustable setting.
- **Conceptual Searching**
  - As noted above, conceptual searching is powered by different types of indices than traditional searching. Conceptual searching uses these indices to try to return results based on related ideas and topics rather than just based on whether the same specific words and phrases are used.
- **Other Tools and Features**
  - In addition to these core search functions, most review tools also offer a range of reporting and administration tools (e.g., saved searches, search history, etc.) to assist you in brainstorming, testing, and iteratively improving searches to meet your information needs. Many tools now also offer some form of word cloud or topical heat map feature to facilitate visual review of the most used words or phrases in your materials.

5.3.2 Filtering

In addition to your searching options, most platforms also offer you a range of options for sorting and filtering by specific properties of documents to help you surface what matters and prioritize what matters most. Most often this is based on a combination of metadata values extracted from the documents, such as file type and date, and custom-created metadata values, such as domain name or custodian.

Often, these types of sorting and filtering capabilities are now tied to visualization tools that let you see the distribution of materials (and any gaps in them) at a glance and that allow you to adjust a range of value limits to see how they narrow or expand your results. For example, many tools now offer communication maps that can show which people are communicating with each other, how often they are doing so, and other useful details.

5.3.3 Searching and Filtering and the Three Goals

Just as we noted that sampling is excellent for finding *unknown* unknowns (the materials for which you don't know you need to be looking), so searching is excellent for finding your known knowns and unknowns (the materials for which you do know you need to be looking). Targeted searching of key words, names, and phrases is one of the fastest ways to find relevant materials and hot documents. It is where practitioners focused on the traditional ECA goal typically start.

Filtering and visualization tools, on the other hand, are excellent at EDA – helping you assess the completeness of your collection, the most important dates and sources, the connections between custodians, and more. Leveraging these tools can quickly identify gaps that need to be filled through further collection and rich veins of materials that should be prioritized for further assessment. They can also assist in traditional ECA by illustrating the flow of communication between and with the key players.

Both searching and filtering are important to the goal of Downstream Prep. When looking ahead to review and production, your goal is to eliminate as much of the chaff as possible without losing

an unreasonable amount of the wheat. The more you can refine your search or searches, the more searches you can apply, and the more filters and exclusions you can apply, the more of that chaff you can eliminate, thereby reducing the cost and duration of all downstream activities. Even when collection has been narrowly targeted and not much material can be eliminated during ECA, using these tools and techniques to prioritize and organize your materials will still yield savings and quality benefits in downstream activities.

5.4 THREADING, DUPLICATES & NEAR-DUPLICATES

After sampling tools and searching and filtering tools, the next major types of tools available for pursuing the three goals of ECA are tools for handling email threading and for handling duplicates and near-duplicates.

5.4.1 Threading

Despite the rise of mobile and social sources, collaboration tools, and other alternative communication channels, email still remains a major component of most ESI collections for eDiscovery, and it tends to be voluminous. A single gigabyte of email can easily contain 5,000 to 10,000 discrete email messages, plus their attachments and embedded images and objects. Thankfully, email ESI also typically contains a significant amount of repetition and overlap that can be skipped.

For example, if you collect email from two custodians, you will have multiple copies of the email messages sent between them – a sender copy and a recipient copy for each one. Moreover, if they are engaged in a thread of replies to each other, the emails in such a thread may contain the preceding emails within themselves as quoted text, and the last one in the thread may contain the full text of the whole thread within itself. Such emails are sometimes referred to as inclusive emails, as are any standalone or offshoot emails that contain unique content or attachments.

Email threading tools typically offer some version of two functions to users: conversation threading and inclusive email identification. Additionally, as noted above, many now offer visualization features as an alternative way to explore the email threads and inclusive emails identified by the system.

- **Conversation Threading**
  - Conversation threading is a process in which emails are analyzed and automatically organized into thread groups, arranged chronologically. This analysis looks at existing conversation IDs, if available, and a range of email header fields and other document properties to match up replies in sequence. Such organization makes it possible to quickly identify related materials, speeding up investigation, and to quickly see the context surrounding a particular message, improving understanding. Additionally, presenting emails to reviewers as organized threads speeds up later document review.
- **Inclusive Email Identification**
  - Inclusive email identification is a process in which textual analysis is used to identify inclusive emails, i.e. those that contain a full thread within themselves or that otherwise contain unique text or attachments. Identifying the inclusive emails allows you to more quickly get the full picture, speeding up investigation, and when used as a filter, it can dramatically reduce the number of emails requiring later document review.

5.4.2 Duplicates

The operation of computer systems can produce a lot of duplicate files (including duplicate emails, as noted above). Although duplicates may need to be tracked and reported on in certain circumstances, they do not need to be examined during ECA or included in later review. Such duplicate files are identified using a technique called hashing.

Hashing is a technique by which sufficiently unique "fingerprints" can be generated for files. Hash functions are mathematical processes that take irregular-length inputs (e.g., the data in a particular file), and use them to generate fixed-length outputs (e.g., a string of 32 numbers and letters). Hashing for duplicate identification is accomplished using a cryptographic hash function (e.g., MD5 or SHA-1), which is well-suited to matching unique inputs to particular outputs. Identical files produce identical hash values, and hash values can be easily compared by software to automatically identify matches across even a large collection of ESI.

- **Duplicate Identification**
  - Typically, collected ESI is hashed and deduplicated during, prior to the ECA phase of the project. But, because other rules (e.g., family group preservation) may override deduplication, some duplicates may remain. For example, if the same spreadsheet was attached to two different emails, neither copy would be removed. Most platforms provide features for identifying and managing such duplicates within your loaded collection of materials.
- **Repeated Content Identification**
  - In addition to identifying fully-duplicated documents, many platforms also offer some form of repeated content identification. Such features are designed to automatically identify frequently-repeated blocks of text (e.g., email signature blocks, automatic confidentiality warnings, etc.) so that they can be filtered out of search results (reducing false positives, particularly for privilege searches) and omitted from the creation of semantic/conceptual indices (improving the effectiveness of the semantic/conceptual analytic tools we will discuss below).

5.4.3 Near-Duplicates

In addition to true duplicates, it is common for collections of ESI to contain large numbers of near-duplicates. Near-duplicates are documents that are substantially similar to each other, but not truly identical (and therefore not removed during deduplication). There are two main types of near-duplicates that occur:

1. Superficially-identical documents that only vary in some metadata property, typically arising from their different sources or collection methods
2. Documents with some actual variation in content, like successive drafts of a contract

Finding the former reduces the number of documents to consider (and later review), while ensuring consistent treatment across duplicates. Finding the latter can provide valuable context to the development of key documents over time.

- **Near-Duplicate Identification**
  - Textual near-duplicate identification is somewhat more complicated behind the scenes than true-duplicate identification. Rather than comparing whole documents as single, abstracted values, the full textual content of the documents must be broken down into smaller pieces (sometimes called shingles). These small pieces can then be hashed and the sequences of those pieces compared across documents. If a sufficient number of pieces match, in the right order, the documents will be treated as near-duplicates. Typically, the threshold of similarity at which the system treats two documents as near-duplicates can be customized.

5.4.4 Threading and Duplicate Tools and the Three Goals

These threading and duplicate tools can yield benefits for each of the three ECA goals:

- **For Traditional ECA**, conversation threading provides quick access to related messages and substantive context, as can near-duplicate identification. The ability to identify and filter repeated content also improves your search and advanced analytics results, making investigation easier.
- **For EDA**, the ability to map conversation threads and inclusive emails can reveal gaps requiring further collection, the identification of excessive duplicates can reveal low-priority tranches of ESI, or the identification of excessive near-duplicates might reveal a collection process issue.
- **For Downstream Prep**, the ability to organize conversation threads together and near-duplicates together can dramatically speed up later review work, and the ability to eliminate non-inclusive emails and true-duplicates from that review work can dramatically reduce volume.

5.5 ADVANCED ANALYTIC TOOLS

After thread and duplicate management tools, the final major tools and techniques available for pursuing the three goals of ECA are advanced analytic tools, powered by semantic indexing and other advanced mathematical analyses, including: concept searching, concept clustering, categorization, and TAR 1.0 and 2.0 workflows.

5.5.1 Semantic Indexing

As we noted briefly above, there are more sophisticated types of indices than the traditional inverted indices used to power basic search functions. These semantic indices analyze the available materials in a different way to power different kinds of features. Whether created by latent semantic analysis or probabilistic latent semantic analysis or another related mathematical approach, these indices are designed to go beyond just listing all of the words in a document to reveal the semantic content of those words.

This semantic analysis is accomplished by analyzing the co-occurrence of unique terms across the collection of documents (e.g., how often does the term "fire" appear with the term "employee" and how often does it appear with the term "extinguisher"). This analysis of co-occurrences is used to create an n-dimensional map (like a traditional map of Cartesian coordinates, but with many more dimensions than just x, y, and z). The more frequently unique terms co-occur together, then the stronger the relationship between them, and the more co-occurring terms in two documents, then the closer to each other they will appear on the map. Dense clusters of such documents suggest key topic areas or concepts in the document collection (e.g., employee termination discussions in one area of the map and fire safety discussions in another).

5.5.2 Features Powered by Semantic Indexing

Semantic indices are used to power a variety of branded features in different eDiscovery software platforms, but regardless of name or variation, there are three key functions that are generally available:

- **Concept Searching**
  - Searching against a semantic index does not require an exact match in the way that searching against an inverted index does. Instead, the terms or phrases you search are mapped onto the existing index and documents that are close enough to those search terms on the map will be returned as results – even if none of the exact terms you searched appear. Some concept searching features are referred to as natural language search features, and some also offer an option to search for more documents like a given example document, which may be a real sample or a synthetic one, created for the purpose.

    Another advantage of searches against these indices is that they can reveal more than just a binary, yes-or-no result. Because of the nuanced multidimensionality of the index, you can get results scored on how responsive or not responsive they are to your search (i.e., how close or far away the result was on the map).

- **Concept Clustering**
  - Concept clustering is an automated, unsupervised process in which software analyzes the semantic index that has been created. Rather than looking for the closest matches to a user-provided search, the software looks for the densest clusters of related materials it has identified and groups those results together into clusters defined by their most frequently occurring terms. How dense a cluster must be to qualify is typically a customizable property. Those clusters can then provide an alternative way to explore a collection of documents, to learn about the scope of topics and range of materials it contains, and to identify areas for further exploration.
- **Categorization**
  - Categorization is akin to a hybrid between concept searching and concept clustering. It is a process in which a user selects a set of example documents to define a cluster for the software, and then the software attempts to find all the other documents that should go in that cluster with the examples provided. This is one of the basic workflows underlying technology-assisted review – or what we will refer to as TAR 1.0.

5.5.3 Technology-Assisted Review

Technology-assisted review is used to refer to a family of workflows that leverage categorization (or similar functions), in combination with sampling, to achieve a reliable document review process that requires significantly fewer hours of manual, human review than traditional all-manual approaches. Since its initial rise to prominence in 2011, the available array of TAR tools has expanded and evolved, and eDiscovery service providers have continued to develop new workflows to leverage them in useful ways for the diverse range of projects their clients face.

Although full deployment of a TAR workflow is typically part of the review phase of an eDiscovery project, these workflows – or limited versions of them – may also be leveraged to explore a collection during ECA, to organize and prioritize it for a more traditional review process, or to create a yardstick against which to measure a more traditional review process.

As noted above, TAR approaches come in two major varieties, which we will refer to as TAR 1.0 and TAR 2.0:

- **TAR 1.0 – LSI, Predictive Coding**
  - TAR 1.0 refers to the initial, categorization-based workflows offered in eDiscovery – many of which were, and are, referred to as predictive coding. Broadly speaking, these workflows involve leveraging a sampling process to create a training set or seed set (i.e., a user-defined cluster or clusters), which the chosen software than uses to find other similar documents. These results are then reviewed and coded, and that coding is used to improve the software's results. This training cycle is iterated multiple times until an acceptable quality of results is achieved. The effectiveness of the whole process is measured using either a previously prepared control set or an additional random sampling effort.
- **TAR 2.0 – SVM, Continuous Active Learning**
  - TAR 2.0 refers to more recent workflows developed to leverage new tools based on a mathematical approach called support vector machines (SVM). Rather than being based on identifying the similarities in a large, prepared training set like categorization and TAR 1.0, these workflows are characterized by continuous active learning that updates relevance scoring and prioritization for all documents dynamically as each additional document is coded by a reviewer.

    This is accomplished by focusing on a single, binary classification (i.e., relevant to topic X and not relevant to topic X) and analyzing the differences in language between successive, single example documents to identify the hyperplane that best divides the relevant examples from the non-relevant examples on a multidimensional map. Each additional example the software analyzes and maps can lead the software to identify a more efficient hyperplane between the two groups, improving its classifications.

    These workflows emphasize speed over structure, and so, they work best in situations where there is a clear, binary classification decision to make and where family groups and other contextual factors are less important than overall speed.

5.5.4 Advanced Analytic Tools and the Three Goals

These advanced analytic tools can yield benefits for each of the three ECA goals:

- **For Traditional ECA**, advanced analytic tools are some of the most powerful. Concept searching lets you find relevant materials even before you know the best search terms to use; concept clustering lets you explore a cross-section of topics to find unknown unknowns and identify areas for further exploration; and, categorization can let you use a few relevant examples – including synthetic ones you make – to find more. And, if you are dealing with a time-sensitive, investigative matter, TAR 2.0 workflows may be able to rapidly surface relevant materials.
- **For EDA**, concept clustering can provide a valuable overview of your materials – including revealing an absence of things you expected or the presence of things you don't need, which can inform decisions about ongoing collection activities.
- **For Downstream Prep**, concept clustering can help organize and prioritize subsequent review activity – including both areas to review first and irrelevant areas to skip (e.g., travel emails, fantasy football emails, etc.). Employing categorization or a TAR workflow of some kind can also be used for the same purpose, or to create a yardstick against which subsequent manual efforts can be measured.

5.6 ALIGNING YOUR LENSES TO SEE THROUGH THE FOG

Our survey of tools and techniques for early case assessment has revealed a wide range of available options, each with different strengths and intended applications, but achieving effective ECA is not a question of applying as many of these tools and techniques as you can. Rather, it is a question of selecting the right ones to best serve your primary goal – whether that's Traditional ECA, EDA, or Downstream Prep – and then building on those initial steps in a rational way to eventually achieve all three goals over the course of your ECA efforts.

To return to our earlier analogies, it is a question of aligning the right lenses, in the right order, to peer through the fog of war and bring your informational quarry into sharp focus.

5.6.1 For Pursuing Traditional ECA

When your top priority is pursuing the Traditional ECA goal, the first question to ask yourself is how much knowledge you have of what you expect to find. If you know a lot about what you're looking for in your ESI (e.g., from thorough custodian interviews, from overlap with prior legal matters, etc.), you may be able to jump right to searching for it. If you don't know a lot about the materials you're seeking, which is more common, you will want to start with one or more of the tools and techniques best suited to revealing unknown unknowns:

- Formal random sampling to estimate prevalence, which lets you see a cross-section of everything you have and some of all the different terms and phrases your custodians use to help you better plan your next ECA steps
- Visualization tools (e.g., communication maps, word clouds, etc.), which can reveal patterns of communication and behavior and assist with completing the picture of what happened in other ways

- Semantic indexing features, which let you use concept searching to find relevant materials without knowing the best search terms, concept clustering to explore a cross-section of topics, and categorization to use a few relevant examples to find more
- TAR 2.0 workflows, which can rapidly surface relevant materials in certain circumstances

Once you start to get a handle on what you are really seeking (or if you already knew), you can transition from these initial, exploratory efforts to more targeted search and filtering efforts, which can quickly find relevant materials and hot documents. And, as you find relevant materials to review, thread and duplicate management tools can be used to find related materials to review for context as needed (e.g., related emails, alternate drafts, etc.).

5.6.2 For Pursuing EDA

If your top priority is pursuing the EDA goal, finding individual documents and facts is less important than ensuring sufficiently complete collection has taken place and that any filtering applied during processing has not been excessive. In such situations, your focus should be on tools and techniques that help you see the big picture of your ESI collection and reveal the gaps within it:

- Metadata filtering and visualization tools, which help you assess the completeness of your collection by revealing ranges of values and gaps in those ranges, as well as potentially revealing important date ranges and sources, the connections between custodians, and more
- Concept clustering, which can provide a valuable overview of the content types and topics within your materials, including revealing an absence of things you expected or the presence of things you don't need
- Visualization tools (e.g., communication maps, word clouds, etc.), which can reveal collection gaps, including missing date ranges, missing custodians, and more
- Thread and duplicate management tools, which can provide another way to map conversation threads to reveal gaps requiring further collection, or which can reveal the presence excessive near-duplicates suggesting a collection or processing issue

Formal random sampling can also be useful during EDA, particularly if there are disputes over the appropriate scope of preservation and collection that need to be resolved. Sampling to estimate prevalence can be used to apply relative value determinations to different sources and tranches and to estimate costs and benefits associated with specific proposed work.

5.6.3 For Pursuing Downstream Prep

When your top priority is pursuing the Downstream Prep goal, you are concerned with learning about what happened, but only insofar as that informs what must be reviewed later and how it should be prioritized. And, you are concerned with understanding the properties and the big picture of the ESI you've collected, but only insofar as that informs what tools and techniques for culling you should choose and what review methodologies are likely to be effective. All of the tools and techniques discussed so far can be leveraged to assist in the Downstream Prep effort:

- Formal random sampling to estimate prevalence, which allows you to accurately estimate what you have, to evaluate the suitability of potential review workflows (including assessing the viability of a TAR or CAL solution or the need for additional

objective culling), and to create a yardstick against which to measure future review work

- Formal random sampling to test classifiers, which allows you to iteratively improve any searches you plan to apply for culling, to ensure that they minimize unnecessary downstream review work and that they avoid missing any important materials
- Searching and metadata filtering, which can both be leveraged to eliminate as much of the chaff as possible without losing an unreasonable amount of the wheat, thereby reducing all downstream review and production costs
- Thread and duplicate management tools, which can dramatically speed up later review work, both by eliminating materials not requiring review and by providing superior organization to what remains
- Semantic indexing features, which can let you use concept clustering to help organize and prioritize subsequent review activity or let you leverage TAR workflows

5.7 KEY TAKEAWAYS

There are seven key takeaways from this chapter to remember:

1. Early case assessment has evolved to encompass three connected-but-distinct goals: traditional early case assessment, early data assessment, and downstream activity preparation.
2. To pursue the three goals, ECA and review software platforms include a range of useful features, including: sampling tools, search and filtering tools, threading and duplicate tools, and advanced analytic tools.
3. Sampling tools and techniques – particularly formal sampling – are good at revealing unknown unknowns, at giving an accurate overview of your materials, and at providing a reliable basis for improving searches and planning subsequent steps.
4. Search and filtering tools, including newer visualization tools like communication maps, are good at finding specific materials, at identifying gaps in your collection, and at eliminating irrelevant materials prior to review.
5. Threading and duplicate tools are good at placing documents in context, at prioritizing and organizing materials for review, and at avoiding duplicative review.
6. Advanced analytic tools are good at exploring a collection of materials without foreknowledge of the contents and key terms, at rapidly surfacing relevant materials, and at prioritizing and organizing materials for review.
7. Successfully achieving your ECA goals requires leveraging the right combination of these tools based on the relative priority of the three goals for your matter, how much prior knowledge of the materials you have, and the amount of time and resources available.

Chapter 6

The Main Event: Review Fundamentals

About this Chapter

In this chapter, we will review the fundamentals that practitioners need to know about document review, including: what gets reviewed; for what it gets reviewed; by whom it gets reviewed; workflow design considerations; and finally, quality control.

6.1 THE MAIN EVENT

Document review has long been the most expensive and time-consuming phase of a discovery project. One oft-cited study from 2012 estimated[1] that **73%** of total discovery costs were attributable to document review. Even with today's more sophisticated tools and techniques, document review remains the largest discovery expense in a typical matter. A more recent survey found that, in 2023, document review still accounted for **65%** of discovery expenditures.[2]

The reason for these significant costs is the need for qualified people to spend time looking at a significant number of documents to make nuanced determinations about their relevance, their privilege, and more. Regardless of whether that work is concentrated in a small case team, spread among a large contract review team, or outsourced to a managed review service, the total time required has the potential to be in the hundreds or thousands of hours.

Moreover, the quality and consistency of all those hours of work must be ensured so that all relevant information can be uncovered, so that complete responses can be provided, and so that inadvertent disclosure of privileged or confidential material can be avoided.

6.2 WHAT GETS REVIEWED

The first step in any review project is determining what materials are going to get reviewed, because the volume and composition of those materials will inform your subsequent decisions about who does the reviewing and how they go about it. Without a clear picture of the what, you cannot make an effective plan for the who and the how.

In this context, we are not talking about the kinds of source identification that take place during preservation and collection. Rather, we are talking about identifying what requires review from within the pool of collected, processed materials already loaded into an eDiscovery platform. Traditionally, during processing, this pool of loaded materials has already:

- Had its system files, etc., removed
- Had its duplicates identified and removed
- Had any date restrictions applied

This will have left you with a pool of unique files, from the relevant time period, that could contain relevant information. Ideally, you will then have engaged in some early case assessment (ECA) activities to gather information about the contents of this pool to help you decide what gets reviewed and how best to go about it.

From this pool then, you must decide whether everything gets reviewed, whether only the results of certain searches and filters get reviewed, whether only the results of a technology-assisted review (TAR) process get reviewed, or whether some hybrid plan is employed:

- **Everything**
  - For smaller pools of materials (i.e., those containing only a few thousand documents) the simplest, fastest solution is often to just review everything.

[1]Nicholas M. Pace & Laura Zakaras, *The Cost of Producing Electronic Documents in Civil Lawsuits: Can They Be Sharply Reduced Without Sacrificing Quality?*, RAND Corporation (2012), available at https://www.rand.org/pubs/research_briefs/RB9650/index1.html.
[2]ComplexDiscovery Staff, *A Current Look at eDiscovery Review: Task, Spend, and Cost Data Points, ComplexDiscovery* (Feb. 9, 2024), available at https://complexdiscovery.com/a-2023-look-at-ediscovery-review-task-spend-and-cost-data-points/.

Reviewing everything is also the typical approach when reviewing productions received from other parties, which can contain thousands of documents.

- **Search and Filter Results**
  - Identifying your ultimate review set through the application of searches and filters (whether established by negotiation or developed during ECA) is the most common approach. This typically requires reviewing the results of the chosen searches and filters, as well as some of the remainder to verify its irrelevance.
- **TAR Process Results**
  - For larger pools of materials, a TAR process may be employed to identify the relevant materials within the pool, or if speed is of the essence, a continuous active learning (CAL) process may be employed to identify the right materials. In some cases, it may also be possible to leverage generative AI tools for such a classification process. Regardless of the assisting software, the identified materials can then be reviewed, along with some of the remainder to verify its irrelevance.
- **Hybrid Plans**
  - It is also common to employ a hybrid of these approaches specific to the exigencies of the case. For example, you might review all of the materials collected from the most critical custodian and then apply searches or a TAR process to the remaining materials. Similarly, targeted searches might be used to quickly identify the most important materials for immediate review, and then a TAR process might be applied to all the lower priority materials afterwards.

Whichever path you choose, you will also need to make decisions about the handling of families, threads, and near-duplicates to finalize your review set:

- **Families**
  - "Families" refers to the family groups of related documents, such as "parent" emails and "child" attachments. If you are reviewing everything, all family group members will already be included in your review set, but if you have applied searches or a TAR process, the results of those efforts may not be family group complete. You will have a choice about whether to pull related family members in too, or to just review the actual results. Most of the time, they are included – both for the context they provide and because production in complete family groups is common.
- **Threads**
  - "Threads" refers to the threads of related emails going back and forth between participants, which often contain within themselves the text of the messages that preceded them. The single email at the end of the thread may contain the complete thread within itself. Such emails are called "inclusive emails." Most review platforms will give you the option to identify inclusive emails and limit review to just those, excluding from the review set all of the individual preceding emails (and some allow review decisions to be propagated across a thread).

- **Near-Duplicates**
  - "Near-duplicates" refers to those documents that are extremely similar (or superficially identical) to other documents in your collection but that were not removed by deduplication during processing due to some small variation(s) between them (e.g., edits in successive drafts, differences in metadata values). Most review platforms will also give you the option to identify near-duplicates, either for grouped inclusion in the review set, or to exclude all but one instance (and then propagate review decisions across the group).

6.2.1 Scope and Process Negotiations

One of the most important factors in determining what gets included in your review set is the scope limitations and process decisions you negotiate with the other parties before, during, and after the meet and confer. It is common to negotiate agreements to limit the scope to specific custodians, to specific enterprise sources, to specific date ranges, to specific file types, and more. It is also common to negotiate over what searches should be run, what TAR process should be used, and other aspects of the review set identification process. The more scope limitations you can negotiate, the less time and money you will have to spend on review, and the more process elements you can negotiate up front, the fewer decisions you may need to defend later.

6.3 FOR WHAT IT GETS REVIEWED

The next aspect of review to consider is for what your identified review set needs to be reviewed, including:

- Relevance and responsiveness
- Privilege
- Confidentiality
- Deposition preparation

6.3.1 Relevance and Responsiveness

When planning and executing a document review effort, it is important to remember that relevance and responsiveness are distinct things:

- Relevance, as defined by Federal Rule of Evidence 401,[3] is a question of whether a particular piece of evidence "has any tendency to make a fact more or less probable than it would be without the evidence" and "the fact is of consequence in determining the action." And discoverability, as defined by Federal Rule of Civil Procedure 26,[4] extends to any evidence that is both relevant and proportional.
- Responsiveness, in contrast, refers to whether or not a given piece of evidence is responsive to any proportional discovery request propounded by another party. The universe of responsive materials should be a subset of the universe of relevant materials.

[3]Fed. R. Evid. 401, available at https://www.law.cornell.edu/rules/fre/rule_401.
[4]Fed. R. Civ. P. 26, available at https://www.law.cornell.edu/rules/frcp/rule_26.

Everything that is relevant may be helpful to you in understanding the underlying events, and you may wish to plan and execute your review with the intent of finding it all. On the other hand, you might wish to focus your review more narrowly on just finding all the materials responsive to the actual discovery requests received. It is also common to conduct review as a hybrid of these two approaches: applying a top-level tag for relevance versus non-relevance, while also applying request-specific tags to relevant materials that are responsive to one or more specific discovery requests.

6.3.2 Privilege

Reviewing for privilege is of equal importance to finding the relevant and responsive materials within your review set – both because attorneys have an ethical duty to protect client confidentiality (see, e.g., ABA Model Rule of Professional Conduct 1.6[5]) and because inadvertent disclosures can lead to privilege waiver if reasonable steps to prevent the disclosure weren't taken (see Federal Rule of Evidence 502(b)[6]). In addition to the standard attorney-client privilege and work product immunity, you may need to review for other privileges, such as the joint-defense privilege or the physician-patient privilege, depending on the case.

6.3.3 Confidentiality

In addition to privilege, you may also need to review for certain types of confidential information. For example, disclosure of personally-identifiable medical information generally needs to be prevented to comply with HIPAA's Privacy Rule.[7] If you are producing to a federal government agency, you may need to produce a second copy of your materials with confidential business information redacted to prevent disclosure of that information to others through FOIA requests.[8] If you are producing materials collected from within the EU, disclosure of personally-identifiable information may need to be prevented to comply with the GDPR.[9]

Additionally, it is common to negotiate a protective order allowing for the redaction of certain confidential personal information (e.g., phone numbers and email addresses for individual employees) or for the special handling of certain confidential business information to limit who can see it (e.g., trade secrets). Materials subject to such an order will also need to be identified during review.

6.3.4 Deposition Preparation

Later in the discovery process, you may also be reviewing documents – both your own and those produced by other parties – to prepare for depositions. Document review for deposition preparation is different from document review for production. In this context, you are generally re-reviewing materials that have already been determined to be relevant, non-privileged, etc., and you are reviewing them in more detail to create a physical or virtual "witness binder." Such binders may include a chronology, lists of key topics and details, potential exhibits, and more.

[5]ABA Model Rules of Prof'l Conduct R. 1.6 (2021), available at https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_6_confidentiality_of_information/.
[6]Fed. R. Evid. 502(b), available at https://www.law.cornell.edu/rules/fre/rule_502.
[7]U.S. Dep't of Health & Hum. Servs., *Summary of the HIPAA Privacy Rule*, HHS.gov, https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html (July 26, 2013).
[8]U.S. Dep't of Justice, *What is FOIA?*, FOIA.gov, https://www.foia.gov/about.html (Mar. 14, 2011).
[9]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 59, 1 (May 4, 2016), available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504&qid=1532348683434.

6.4 WHO DOES THE REVIEWING

Now that you have made decisions about what you are going to review and for what you are going to review it, you need to make decisions about who is going to perform that review. Broadly speaking, your choices are internal resources (i.e., the case team, existing corporate or firm staff) and external resources (i.e., contract reviewers, managed review services).

6.4.1 Internal Resources

For smaller discovery efforts, it is common for the case team to do most or all of the review themselves, or to do it themselves in conjunction with help from additional paralegals or attorneys already on staff inside the organization or at the primary outside law firm for the matter. Case team members working on the review have the advantage of direct knowledge of the overall matter, its legal issues, etc., and existing staff have the advantages of having already been evaluated as effective team members and of already knowing the organization.

On the other hand, it may eat up a significant amount of the case team members' time engaging directly in review and review management – time that may be more costly per hour than external resources would be. Additionally, it can be disruptive or infeasible to tie up multiple existing employees, for an extended period of time to conduct review, and experienced team members may still be inexperienced document reviewers unable to effectively leverage review tool efficiencies.

6.4.2 External Resources

For larger discovery efforts, some form of external review resources often need to be utilized to supplement, or substitute for, the internal review resources described above. Broadly speaking, external review resources come in two types: contract review staff and managed review services:

1. **Contract Review Staff** – A variety of discovery services providers and staffing agencies provide document review attorneys on a contract basis, at an hourly rate. The hiring organization or law firm can typically specify required experience levels, required language skills, required knowledge (e.g., a chemistry background), and more. This can facilitate supplementation of an internal team for scale or specialization. Once hired, however, the hiring organization or firm is then responsible for providing these contract reviewers with secure space, with workstations and systems access, and with assignments and oversight. You are being provided with reviewers rather than with review, which limits the scalability of this approach.
2. **Managed Review Services** – Managed review services, on the other hand, provide review rather than reviewers. Such services, whether onshore or offshore, maintain their own pools of reviewers and review managers, usually a mixture of permanent

staff and experienced, pre-vetted contract review staff. They also maintain their own secure environments, as well as standardized review, quality control, and documentation processes. Case teams still dictate review goals, assist in review team training, resolve review questions as needed, and evaluate review results, but most of the actual review and the management of the review are handled by the service provider.

6.4.3 Reviewer Training

Whether your team is internal only, internal plus contract, or entirely external, it is important that the reviewers have a clear and consistent understanding of what things they are looking for, what standards they are applying, and what processes they are following. For example:

- What is the scope of relevance for the case?
  - What are the meanings of any specific requests?
  - What qualifies as a "hot" document?
- What context do they need to know?
  - About the organization?
  - About the underlying events?
  - About the primary legal issues?
- Are they checking for privileges?
  - Which ones?
  - Using what standards?
  - What about HIPAA, CBI, PII, etc.?

It is common to provide review teams with a written review protocol document that provides answers to all of these questions, along with relevant background information and example documents from the collection. This protocol and the associated examples are typically reviewed with the team during an initial training and question session, and then follow-up questions are addressed by the case team as needed throughout the review.

6.5 WORKFLOW DESIGN CONSIDERATIONS

Once you know what materials you're reviewing, for what properties you're reviewing them, and who's doing that reviewing, you can plan the actual workflow by which the review work will be executed. If you are designing the review workflow yourself, rather than relying on a managed review service, you will need to consider document flow, tagging palettes, batch creation, and process documentation.

6.5.1 Document Flow Considerations

Designing an effective document review workflow is a project-specific exercise that requires consideration of a wide range of options and factors, including: the features and functions available to you in your chosen document review platform, the volumes and types of materials being reviewed, the number and nuance of things for which the materials must be reviewed, the

number and skill level of the chosen reviewers, and the available time for completion of the review.

Smaller, simpler projects may require only a simple workflow, with just a traditional first level review checking for both relevance and privilege, and a second level quality control review double-checking some of that work prior to production. More complex projects may call for multi-level, multi-path workflows with specialized teams handling specific tasks. For example:

- Projects with numerous, nuanced responsiveness determinations to make might call for separating initial relevance review from subsequent issue responsiveness coding. Each additional determination a reviewer must make on a document decreases their review speed, and having too many determinations to make will increase their error rate.
- Projects with high volumes or with nuanced privilege issues might call for separating privilege review from relevance/responsiveness review, having it performed by particularly skilled reviewers only for the materials deemed responsive.
- Projects with a high volume of materials requiring redaction (for privilege, confidentiality, etc.) may separate redaction into its own step, handled by a dedicated team, rather than asking the first-level reviewers to complete redactions.

6.5.2 Tagging Palette Considerations

As we noted above, there is a tension in document review between speed, accuracy, and nuance: the more determinations a reviewer must make, the longer it will take them, and the more mistakes they will make. Understanding this tension is important when creating the tagging palette your reviewers will use to annotate documents with their determinations.

Reviewers only working with tags for simple relevance, potential privilege, and hot documents will be able to work more quickly and consistently than those who must also apply tags for specific issues, specific privilege types, and other nuances. A good rule of thumb is to try to keep each reviewer from having to make more than five determinations at a time about each document. If many more than that are required, consider breaking those determinations up across multiple review passes or paths. Some platforms allow for the creation of multiple, separate tagging palettes to support complex workflows involving multiple teams.

Depending on your workflow and your chosen platform's built-in review tracking features, you may also need to include tags designed to aid you in:

- Tracking documents' progress through your workflow
- Tracking who's reviewed them at each step in the workflow
- Tracking whether tagging changes have been made during quality control

Ideally, you should rely as much as possible on the review tracking functions built into your chosen platform to minimize complexity in the tagging palette(s) being used.

6.5.3 Batch Creation Considerations

In addition to planning your document flow and creating your tagging palette(s), you will also need to make some decisions about how the large pool of documents to be reviewed should be broken up into batches for reviewers to complete:

- **How should your review pool be organized into batches?**
  - Depending on your review goals and priorities, you might break up your review pool into batches by custodian, by search term hits, by concept clusters, by chronology, by source type (e.g., batching text messages together, emails together, etc.), or by other factors.
- **How should threads and near-duplicates be handled?**
  - As we discussed above, you will need to decide whether you are including or excluding near-duplicates and non-inclusive emails and, if so, you will need to decide whether to keep them grouped together during batch creation.
- **How should family groups be handled?**
  - As we also discussed above, you will need to decide whether you are keeping family groups of related records together; if you are planning to produce in complete family groups (most common), it is generally best to create review batches that way too, both for the additional context it provides, and so that all family members get reviewed prior to production.
- **How should each batch be sorted?**
  - You may also be able to specify the default sorting for the materials within each batch; sorting them chronologically is the most common choice (this can often be done by a family group master date rather than each document's individual date to maintain family groupings together within the chronological sort).
- **How large should each batch be?**
  - Batch size should be selected based on how you want your reviewers to work; it is generally best to keep batch sizes small enough that they can be completed in 1-2 hours, as error rate increases the longer reviewers go without a break; how many documents that is will depend on your documents, but batches of 50-100 documents are common.

Another factor that can affect the speed of your reviewers' work is the mix of file types and file lengths that they receive in each batch of documents they review. While the majority of documents are likely to be text documents of moderate length through which they can move at a quick, even pace (e.g., emails and Word documents), some may be outliers that will break the rhythm of their work, such as:

- Multimedia files requiring a switch to listening or watching
- Large spreadsheets requiring a switch to native review
- Very long documents requiring protracted reading time

If you are running a large, time-sensitive review, it may well be worth the effort to preemptively filter such files out of the general review pool before batch creation (by file type, file size, etc.). Once segregated, those rhythm-breakers can be grouped into their own batches, by type, for separate review.

6.5.4 Documentation Considerations

When engaged in design of a review workflow, you will also need to think in advance about the documentation needs you will have during the course of the review. Generally, you will want some way to track:

- Your overall progress, your progress against budget, and your rate of progression
  - To project remaining time and cost to completion
- Your rates of relevance, privilege, redaction needed, etc.
  - To project the production, privilege logging, and redaction work still to be done
- The speed and accuracy rates of individual reviewers
  - To identify and address misunderstandings and performance issues

Additional metrics may be also tracked for both intra- and inter-project benefits.

Once you're tracking your chosen metrics, you will also need to generate reports to share and contextualize the important information with relevant team members, client representatives, etc. Frequency and content is entirely dependent on your needs, but it is common to provide weekly review progress reports, often with some additional reporting done monthly. Although all of this tracking and reporting can be done manually, most review platforms now include features to address these needs.

In addition to tracking and reporting on aspects of your project's progress, you will also want a plan for documenting decisions about the review project. In the event that there is a later challenge to your methods and their results, it will be invaluable to have contemporaneous notes or emails documenting why you did what you did the way you did it – both as potential evidence and to refresh your recollection of decisions made months or years before.

6.6 QUALITY CONTROL FUNDAMENTALS

The final and most important fundamental of review to understand is quality control. No matter what you're reviewing, what you're reviewing it for, who's reviewing it, or how you're reviewing it, you will need to take proactive steps to ensure the overall quality and consistency of that work. Perfection isn't possible and isn't[10] required,[11] but reasonable efforts to meet your obligations of completeness, accuracy, and privilege protection are both.

6.6.1 The Myth of the Gold Standard

The Sedona Conference's Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery[12] describes a persistent myth in eDiscovery:

> It is not possible to discuss this issue without noting that there appears to be a myth that manual review by humans of large amounts of information is as accurate and complete as possible – perhaps even perfect – and constitutes the gold standard by which all searches should be measured.

[10] *See, e.g., Dynamo Holdings Ltd. P'ship, et al., v. Comm'r of Internal Revenue,* 143 T.C. No. 9 (USTC Sep. 17, 2014), available at https://www.millerchevalier.com/sites/default/files/resources/Dynamo_USTaxCourt.pdf.
[11] *See, e.g., Winfield v. City of New York,* 2017 WL 5664852 (S.D.N.Y. Nov. 27, 2017), available at https://docs.justia.com/cases/federal/district-courts/new-york/nysdce/1:2015cv05236/444418/217.
[12] The Sedona Conference, *Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery,* 15 SEDONA CONF. J. 217 (2014), available at https://thesedonaconference.org/publication/Commentary_on_Search_and_Retrieval_Methods

The reality is quite different from this myth. In reality, even the best reviewers make mistakes due to simple human fallibility, and reviewers frequently come to different conclusions regarding questions of relevance, privilege, and more. Studies have shown[13] surprisingly low consistency between the independent results of equivalent review teams ("Assessor Overlap").

Because of this reality, it is critical that your document review project include some steps to ensure an acceptable minimum level of quality, consistency, and completeness.

6.6.2 Traditional Methods

The most traditional method of quality control is second level (or second pass) review. In this method, some portion of the material reviewed by first level (or first pass) review is re-reviewed by more senior reviewers to check the accuracy and consistency of the work. The volume re-reviewed and the focus can vary widely depending on the needs of the project:

- In a smaller project, you might re-review everything deemed relevant and non-privileged to make sure nothing irrelevant or privileged is produced.
- In a larger project, you might re-review a random 10% of the first level review to look for recurring mistakes to address, or problem reviewers to retrain or replace.
- In a project using a TAR, CAL, or AI workflow, you might focus more on evaluating the materials deemed irrelevant to be sure nothing important has been missed.

In some projects you may establish more than two levels of review. For example, you might add a third level in which case team members re-review certain materials prior to production.

The other traditional quality control method is targeted searching. Targeted searching is the practice of running searches against the reviewed materials for key terms that would likely indicate clear relevance, irrelevance, or privilege and then double-checking that the results are coded correctly. For example, you might search for key attorneys' names and email addresses and then double-check the privilege tagging applied to the results.

6.6.3 Sampling

Sampling comes in two broad categories: judgmental sampling and formal sampling. Judgmental sampling is the informal process of looking at some randomly selected materials to get an anecdotal sense of what they contain. The random 10% second-level review and targeted searching described above are examples of judgmental sampling. The goal of these efforts is to get an impression and make an intuitive assessment rather than to take a specific measurement.

Formal sampling is just the opposite: you are reviewing a specified number of randomly-selected documents with the goal of taking a defined measurement with a particular strength. Typically that measurement is either being taken to test classifiers or estimate prevalence:

- **Testing Classifiers**
  - This is the process of seeing how effective and efficient a particular classifier actually is, be it a search, a TAR process, or a human reviewer. Using this technique, you can quantify the accuracy and error rate of individual reviewers and teams or quantify the recall and precision of searches or TAR processes.
    - In the context of quality control, these measurements can be used to identify problem reviewers, to measure overall review effectiveness, or to

[13] Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review,* XVII RICH. J.L. & TECH. 11 (2011), available at http://jolt.richmond.edu/v17i3/article11.pdf.

implement lot acceptance sampling.

- **Estimating Prevalence**
  - This is the process of reviewing a simple random sample of a given collection of materials to estimate how much of a given kind of thing is present.
    - In the context of quality control, this is used most often to measure how much relevant material may exist in the unreviewed remainder left after applying searches or a TAR process (a.k.a. measuring elusion).
    - You might also use this method to create a yardstick for your review before you begin by estimating in advance how much relevant and privileged material you expect to find.

6.6.4 Feedback Loops

Regardless of the specific quality control methods you choose to employ on your project, it is critical that effective feedback loops are established. In most document review projects, you will be engaged in ongoing quality control throughout first-level review, giving you the opportunity to not just catch and correct errors, but to identify issues and address them with first-level reviewers to improve the rest of their work. Effective feedback loops make this possible.

A feedback loop between the review managers and the reviewers feeds the insights gleaned from quality control efforts back to the reviewers though additional instruction and clarification. For larger projects, it is common to have weekly review team meetings to review issues and answer questions. It is also common to have one-on-one sessions with individual reviewers identified as requiring additional guidance, and it is a good idea to maintain a shared list of reviewer questions and review manager answers for everyone's reference.

A feedback loop between the case team and the review managers enables the review managers to request guidance and clarification as needed and enables the case team to share any evolution in their understanding of the case, as well as any issues they identify during any quality control review they perform.

6.6.5 The Importance of Privilege Protection

It's worth emphasizing the particular importance of engaging in quality control for the purpose of preventing the inadvertent disclosure of privileged materials. As we noted above, Federal Rule of Evidence 502(b)[14] establishes that inadvertent disclosures can lead to privilege waiver if reasonable steps to prevent the disclosure weren't taken.

[14]Fed. R. Evid. 502(b), available at https://www.law.cornell.edu/rules/fre/rule_502.

The Committee's Explanatory Note on Rule of Evidence 502[15] makes clear that "reasonable steps" is a case-by-case determination that can depend on factors such as the total number of documents to be reviewed, the time constraints for production, how records were managed, what tools were used, and more. Consequently, taking steps to ensure the quality of your privilege review approach[16] is at least as important as what approach you take:

> The implementation of the methodology selected should be **tested for quality assurance**; and the party selecting the methodology must be prepared to **explain the rationale** for the method chosen to the court, **demonstrate that it is appropriate** for the task, and **show that it was properly implemented**. [emphasis added]

6.7 KEY TAKEAWAYS

There are six key takeaways from this chapter to remember:

1. Review is typically the largest discovery expense due to the need for qualified people to spend time looking at a significant number of documents to make nuanced determinations about their relevance, their privilege, and more.
2. Typically, you will review either the results of iterated/negotiated searches or the results of a TAR/CAL process, which are: family group complete; restricted to relevant dates; and, have had system files, duplicates, and non-inclusive emails removed.
3. At a minimum, you will review to identify privileged materials and relevant materials, but you may also need to review for: responsiveness to particular requests, the presence of confidential information, or deposition preparation details.
4. Smaller reviews may be performed by the case team, supplemented as needed with existing staff or contract reviewers, while larger reviews often require the aid of a managed review service. Effective review team training is essential regardless.
5. Selection of review methodology – including review workflow, tagging palette(s), and batch creation – is highly matter-specific and depends upon: the features in your platform, the volumes and types of materials, the number and nuance of needed determinations, the size and skill of the team, and the available time for completion.
6. Quality control – including informal and formal sampling, targeted searching, effective feedback loops, and other steps – is essential to ensuring that review is consistent, results are complete, and privilege and confidentiality are protected.

[15] Fed. R. Evid. 502(b), advisory committee's note, available at https://www.law.cornell.edu/rules/fre/rule_502.
[16] *Victor Stanley Inc. v. Creative Pipe Inc.*, 250 F.R.D. 251, 262 (D. Md. 2008), available at https://casetext.com/case/victor-stanley-inc-v-creative-pipe.

Chapter 7

The Final Countdown: Production Fundamentals

About this Chapter

In this chapter, we will discuss the fundamentals about production that all eDiscovery practitioners should know, including: primary production formats, production format specifics, who gets to decide, example disputes, production preparation, and privilege and production logs.

7.1 THE FINAL COUNTDOWN

Production is another discovery activity, like collection and processing, in which technical decisions can have logistical and legal effects. For this reason, it is important for practitioners to understand the fundamentals of production. How materials are produced affects how long they take to prepare and how easily they can be searched, reviewed, and used later in depositions and at trial.

ESI productions may be anything from a few PDF files to a custom-configured online repository. Negotiating production format, including details like whether and what metadata will be provided, can both ensure maximum usability of what you receive and preempt disputes over what you produce and how you produce it. Failure to understand these options and negotiate effectively about them in advance still leads to frequent disputes today.

Beyond simply being important, the ability to successfully prepare and deliver productions of relevant ESI may also be an ethical requirement for attorneys to fulfill their duty of technology competence. For example, the California duty of technology competence for eDiscovery,[1] explicitly names proper production of ESI as one of its nine core requirements: "produce responsive non-privileged ESI in a recognized and appropriate manner."

7.2 PRIMARY PRODUCTION FORMATS

The first production decision that needs to be made is the format or formats in which the relevant, non-privileged documents will be produced. That decision will determine the workflow that follows for actually preparing, validating, and delivering the production. Broadly speaking, there are four primary production formats available: paper, near-paper, native, and near-native.

1. **Paper**

   In a paper production of ESI, the materials to be produced are printed out and produced as paper documents as in a traditional document production. Per-page Bates numbering, other endorsements, and redactions may be applied. While superficially simple, paper productions can still create technical issues, such as how to display review markup in printed Word documents, how to format large spreadsheets to letter-size pages, or how to handle presentation decks with speaker notes and comments.

   Although, technically, paper production of ESI is an option, it is only logistically viable for matters with few documents and only legally viable for matters with a negotiated agreement to such production. Absent such an agreement, paper production of ESI does not meet the format requirements in the Federal Rules of Civil Procedure we will discuss below.

2. **Near-Paper**

   In a near-paper production, the materials to be produced are converted to image files that simulate printed, paper versions of the documents. Each page image can then have per-page Bates numbering, other endorsements, and redactions applied before production. Such image collections are paired with a load file that

[1] The State Bar of California Standing Committee On Professional Responsibility and Conduct, *Formal Opinion No. 2015-193* (June 30, 2015), available at https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL 2015-193 %5B11-0004%5D (06-30-15) - FINAL.pdf.

records document breaks, provides selected metadata for the documents, and includes (or links to) extracted text for searching. This collection of images and related information can be loaded by the recipient into a document review tool like Relativity.

A near-paper production is a popular choice that combines some of the benefits of a paper production (e.g., per-page numbering, redactions) with some of the benefits of a native or near-native production (e.g., associated metadata, searchable text). It also retains some of paper's drawbacks (e.g., questions of what's visible, unsuitable document types) and creates some new technical issues of its own (e.g., time and cost of image creation, reconciling extracted text with image redactions).

3. **Native**

In a native production, materials are produced in their native formats, as they are created and kept in the ordinary course of business. Such native file collections may also be paired with a load file containing extracted metadata, searchable text, and other information for loading into a document review tool like Relativity, although outside software will still have to be launched to open some types of native files.

A native production can generally be prepared with less time and expense than a near-paper production, and native productions eliminate questions of what's visible in printouts or images and of what metadata is included. Native productions are not without drawbacks, however, as per-page numbering, other endorsements, and redactions are not possible. Moreover, some types of materials may not be reasonably usable in their native format (e.g., email databases, chat logs), and you must be very careful to ensure review of all metadata and hidden content before production. There is also some risk of inadvertent alteration of the native files by the recipient during their review (e.g., auto-date fields in documents updating to display the review date).

4. **Near Native**

A near-native production involves the conversion of native files into another electronic format that approximates the native format. For example, native message logs might be unitized by day or conversation and converted into many separate HTML or XML files. Near-native productions are typically paired with a load file containing extracted metadata, searchable text, and other information. This collection of near-native files and related information can be loaded into a document review tool like Relativity.

A near-native production carries similar advantages and drawbacks to a native production. Certain document types can be presented in a more useful way than they can with a true native production, but you also reintroduce questions of what's visible, of how it's presented, and of what metadata is included. Moreover, per-page numbering, other endorsements, and redactions are still generally not possible.

7.2.1 Hybrid Approaches

In reality, most productions today utilize a combination of near-paper, native, and near-native approaches, applying different handling to different document types to maximize later usability. For example, most documents in a production might be produced as near-paper images (to

7.3 PRODUCTION FORMAT SPECIFICS

Beyond just deciding on your optimal combination of paper, near-paper, native, and near-native production options, there are a range of more-specific options for you to consider. Among the most important are options related to load files, metadata, unitization, redactions, numbering and endorsements, and paper integration.

7.3.1 Load Files

As noted above, many productions are accompanied by a load file that contains information about the various documents and images being produced and that makes it possible for those materials to be imported together into a document review platform. These load files can provide links to native or near-native files, to rendered images, and to extracted text files, and they can contain a variety of fields of metadata and other extracted data for each document.

Load files are essentially large spreadsheets themselves, though their specific formatting requirements and applicable field delimiters vary some from system to system. The specific load file format employed is less important than making sure the parties are on the same page about what format will be used to ensure it's something that works for everyone's platforms.

Decisions will also need to be made (or negotiated) regarding what fields the load file should include, how they should be labeled, and what custom fields – if any – should be created. For example, a field might be included documenting the request number(s) in response to which each document is being produced, or a field might indicate the documents to which a protective order applies.

7.3.2 Metadata

Metadata has tremendous value, both as potential evidence (e.g., revealing when and by who something was modified[2]) and as the basis of many filtering, sorting, and searching options within document review tools. Thus, the metadata fields included in productions will have both evidentiary and usability impacts for the recipients. An appropriate eDiscovery expert can assist you with determining what fields are necessary to meet your needs in a specific case.

In addition to figuring out what metadata fields should be included, you may also need to address what names will be used for those fields, what formats will be used for the values in them, and what time zone should be used to normalize dates and times. The creation of custom fields and values may need to be considered too. For example, should there be a master date field? If so, how should the master date for each document be determined?

For a generic example of essential fields, the EDRM organization's model XML load file includes the following standard metadata and extracted data fields[3]:

[2]Mark A. Berman, *Audit Trail 'Meta Data' Leaves Tell-Tale Signs in Medical Malpractice Actions,* NEW YORK LAW JOURNAL, https://www.law.com/newyorklawjournal/2019/05/06/audit-trail-meta-data-leaves-tell-tale-signs-in-medical-malpractice-actions/ (May 6, 2019).
[3]*Production Guide*, EDRM, https://www.edrm.net/resources/frameworks-and-standards/edrm-model/production/ (Nov. 4, 2010).

- **File Elements**
  - FileName, FilePath, FileSize, Hash
- **Metadata Tags – All Documents**
  - Language, StartPage, EndPage, ReviewComment
- **Metadata Tags – Messages**
  - From, To, CC, BCC, Subject, Header, DateSent, DateReceived, HasAttachments, AttachmentCount, Attachment Names, ReadFlag, ImportanceFlag, MessageClass, FlagStatus
- **Metadata Tags – Files**
  - FileName, FileExtension, FileSize, DateCreated, DateAccessed, DateModified, DatePrinted, Title, Subject, Author, Company, Category, Keywords, Comments

7.3.3 Unitization

The proliferation of mobile device sources, social media sources, and collaboration tool sources has made message thread unitization a common question for eDiscovery. These source types frequently include ongoing threads of back-and-forth messages (e.g., text message threads, direct message threads, Slack channel threads, etc.), which can span long periods of time. Although the specifics vary by source, these message threads are often maintained in ongoing logs that are not conducive to efficient review or later use as evidence. Rather than present weeks or months of messages in a single document, it is typical to unitize these logs into separate, shorter documents for review and production.

When doing so, some judgment must be exercised about what size the units should be. Individual messages stripped of thread context are also not ideal (as courts have pointed out[4]), so some middle ground between massive logs and single messages is preferred. It is common to unitize such materials into 24-hour chunks, so that each day's communications become a single document, but other divisions may be rational depending on your materials and case.

This unitization is typically performed during processing, prior to ECA, review, and production, but production implications should be considered when making the determination, as parties can disagree over the best way to unitize and produce such materials.

7.3.4 Redactions

As noted above, primary production format affects your ability to perform redactions within documents. Generally speaking, native and near-native files cannot be effectively redacted, while near-paper and paper productions can. The availability of effective redactions is one of the reasons for the continued popularity of near-paper, image-based productions.

When preparing a production that will involve redactions, you will need to consider how redactions should appear on the page (e.g., blackout, whiteout, pattern-filled), including whether redaction type (privilege, PII, etc.) affects appearance or requires a label. Additionally, if extracted document text is being provided (to facilitate searching), the extracted text for documents bearing redactions will have to be either excluded or replaced. It can be replaced by performing optical character recognition (OCR) on the page images generated after the redactions have been applied.

[4] *See, e.g., Laub v. Horbaczewski,* 331 F.R.D. 516 (C.D. Cal. Apr. 22, 2019) (Magistrate Judge expressing a preference for "aggregated" formats preserving "the integrity of the threads of communication reflected in the text messages"), available at https://casetext.com/case/laub-v-horbaczewski.

7.3.5 Numbering and Endorsements

Also as noted above, primary production format affects your options with regard to numbering and endorsements. For numbering, paper and near-paper productions allow for per-page Bates numbering, while native and near-native formats generally only allow for per-file numbering to be applied. Combination approaches require coordinating per-page numbering for some documents with per-file numbering for others.

For endorsements, confidentiality warnings or protective order language work the same way. Paper and near-paper productions can have consistent endorsements applied in the headers and/or footers of each page, while native and near-native productions cannot. For native and near-native productions (and, often, for near-paper productions too), custom load file fields may be created that document confidential status, protective order applicability, and other endorsement content for each document.

7.3.6 Paper Integration

Another element to consider is how you will handle paper materials collected during discovery along with all of your ESI. Rather than producing such materials in paper format, you have the option of incorporating them into your electronic production. This can be accomplished by scanning the documents into page images, performing OCR to extract the available text for searching, and manually entering relevant "metadata" values (e.g., bibliographic and source information).

7.3.7 What to Request

The number of potential formats and specific options available can make it daunting to figure out what to ask for when you are negotiating production format or specifying the format in a particular discovery request, and there are technical specifications that may be necessary beyond what we have discussed here (e.g., image format, load file format). Many practitioners find it helpful to consult with an appropriate eDiscovery expert who can guide their selections for a particular case, including providing the relevant technical specifications. Some practitioners also turn to publicly-available production protocols from federal[5] agencies[6] as models they can use for their own production planning.

Despite the many variations possible, the following is a good place to start for many cases:

- **Request a hybrid format production, based on file types**
  - Near-paper image format for all suitable file types (e.g., email, Word, PDF)
  - Native format for the unsuitable file types (e.g., spreadsheets, multimedia)
- **Request production in complete family groups for context**
  - If applicable, specify your preferred unitization approach for message threads

[5] Antitrust Division, *DOJ Standard Specifications for Production of ESI*, U.S. DEPT. OF JUSTICE, https://www.justice.gov/atr/case-document/doj-standard-specifications-production-esi (June 30, 2015).
[6] *Data Delivery Standards*, U.S. SEC. AND EXCH. COMM'N, https://www.sec.gov/divisions/enforce/datadeliverystandards.pdf (Dec. 2020).

- **Request the production of extracted text and metadata in an accompanying load file**
  - Specify the categories or specific fields of metadata you want to receive
  - Specify any custom metadata fields you seek, such as request number

7.4 WHO GETS TO DECIDE

Who gets to decide on the formats and options to be used in a particular case? Pursuant to the Federal Rules of Civil Procedure (FRCP), both parties have opportunities for a say in the production format at three different points in the process:

- First, during meet and confer negotiations
- Second, through actual requests and objections
- Third, through motions to compel and protect

7.4.1 Negotiation in FRCP 26

Production format selection for ESI first comes up in FRCP 26(f)[7] as part of the required meet and confer. FRCP 26(f)(1) specifies that "the parties must confer as soon as practicable," and 26(f)(2) specifies that, among other things, the parties must use the conference to "develop a proposed discovery plan." FRCP 26(f)(3) describes what this discovery plan must address, including "(C) any issues about disclosure, discovery, or preservation of electronically stored information, **including the form or forms in which it should be produced**" [emphasis added].

Good faith efforts to fulfill this requirement are expected of the parties. FRCP 37[8] specifies the consequences for failure in this area:

(f) Failure to Participate in Framing a Discovery Plan. If a party or its attorney **fails to participate in good faith** in developing and submitting a proposed discovery plan as required by Rule 26(f), the court may, after giving an opportunity to be heard, **require that party or attorney to pay to any other party the reasonable expenses, including attorney's fees**, caused by the failure. [emphasis added]

Thus, in an ideal case, the parties discuss the available production formats and other options during their initial conference, reach a mutually-acceptable agreement, and document that agreement in a written discovery plan that both parties then follow.

7.4.2 Requests and Objections in FRCP 34

Unfortunately, many cases do not follow that ideal path, and the specifics of production preferences and expectations remain unaddressed until later in the discovery process. In such cases, FRCP 34[9] provides the next set of instructions for who gets to decide on production format.

In cases where a prior agreement has not been negotiated, FRCP 34(b)(1)(C) allows parties requesting production of ESI to "specify the form or forms in which electronically stored

[7]Fed. R. Civ. P. 26(f), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[8]Fed. R. Civ. P. 37, available at https://www.law.cornell.edu/rules/frcp/rule_37.
[9]Fed. R. Civ. P. 34, available at https://www.law.cornell.edu/rules/frcp/rule_34.

information is to be produced." If the responding party does not want to produce in the requested form, FRCP 34(b)(2)(D) allows them to object to the requested form and state their proposed alternative:

> (D) Responding to a Request for Production of Electronically Stored Information. The response may state an **objection to a requested form for producing electronically stored information**. If the responding party objects to a requested form – or if no form was specified in the request – **the party must state the form or forms it intends to use**. [emphasis added]

If a request does not specify a form (and no form was previously negotiated or ordered), FRCP 34(b)(2)(E)(ii) lays out the ESI production format options from which a responding party can choose: "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**." [emphasis added]

This translates to a choice between producing ESI in native format (the "form or forms in which it is ordinarily maintained") or in some other "reasonably usable form or forms," which typically means near-paper or near-native, accompanied by a load file with relevant metadata and searchable text. Additionally, FRCP 34(b)(2)(E)(i) requires that (unless agreed or ordered otherwise) produced ESI must be organized either as it is "kept in the usual course of business" or labeled "to correspond to the categories in the request."

7.4.3 Motions to Compel and Protect in FRCP 37 and 26

In the event, the request and objection process described above leads to an irreconcilable dispute over the appropriate production format, a requesting party's final recourse is to submit a motion to compel the requested discovery, in the requested format, pursuant to FRCP 37(a).[10] Before doing so, however, the requesting party must make a good faith effort to confer with the responding party to resolve the issue: "The motion must include a certification that the movant has in good faith conferred or attempted to confer with the person or party failing to make disclosure or discovery in an effort to obtain it without court action."

In parallel, a responding party can also seek a protective order, pursuant to FRCP 26(c),[11] protecting it from having to provide the requested discovery, in the requested format. FRCP 26(c)(1)(A)-(C) allows that:

> . . . The court may, for good cause, issue an order to protect a party or person from annoyance, embarrassment, oppression, or undue burden or expense, including one or more of the following:
>
> (A) forbidding the disclosure or discovery;
>
> (B) specifying terms, including time and place or the allocation of expenses, for the disclosure or discovery;
>
> (C) prescribing a discovery method other than the one selected by the party seeking discovery;
>
> [emphasis added]

[10] Fed. R. Civ. P. 37(a), available at https://www.law.cornell.edu/rules/frcp/rule_37.
[11] Fed. R. Civ. P. 26(c), available at https://www.law.cornell.edu/rules/frcp/rule_26.

The same limitation on seeking a protective order applies however: the responding party must first make a good faith effort to confer with the requesting party to resolve the issue.

It should also be noted that, depending on the outcome, either type of motion can result in an award of expenses, including fees, pursuant to FRCP 37(a)(5).[12]

7.5 PRODUCTION FORMAT DISPUTES

Now that we have reviewed how the production format selection process is supposed to work under the FRCP, let's take a look at some example cases to see the disputes that arise and how courts are applying those rules in practice.

7.5.1 A Joint Failure

In *Baker v. Santa Clara Univ.*,[13] the plaintiff sought an order compelling production of requested ESI in native format. During discovery, she "served 54 requests for the production of documents" including "a single request that purports to cover the format of production for all documents responsive to the other 53 requests." That request stated:

> With respect to each request, **produce all documents in native format, including electronically stored information, metadata, and all metadata fields**. Do not do anything that strips, removes, changes, limits, or otherwise alters the actual electronically stored information and metadata fields of any document that exists in an electronic format. Ensure that all such evidence remains intact, undisturbed, and is produced with each and every electronic document. [emphasis added]

Despite this request, the defendant produced over 2,500 pages of materials in PDF format without metadata, and the plaintiff moved to compel reproduction of the materials in native format with metadata.

The plaintiff argued, primarily, that "having these documents in native format will allow her to more easily discover if [the defendant] has omitted responsive documents from its production." The defendant argued that it had attempted to meet and confer about ESI production issues "more than a year ago" pursuant to FRCP 26(f), but that the plaintiff's counsel "did not meaningfully engage in the required discussion." It argued that reproducing now in native format would be "time consuming, burdensome, and expensive."

The court concluded that "[n]either party ha[d] complied with the rules and guidelines that govern the production of electronically stored information," the plaintiff having failed to meet and confer and the defendant having failed to properly object and produce in accordance with FRCP 34(b). The court then looked to "the dual requirements of relevance and proportionality" and concluded that:

> Absent a specific, articulable basis for believing [the defendant] has not complied with its discovery obligations, [the plaintiff] does not have a compelling reason for demanding that [the defendant] re-produce its entire responsive document production in native format simply because she might find something missing.

[12] Fed. R. Civ. P. 37(a)(5), available at https://www.law.cornell.edu/rules/frcp/rule_37.
[13] *Baker v. Santa Clara Univ.*, 2018 WL 3629838 (N.D. Cal. July 31, 2018).

7.5.2 A Protocol Deviation

In *In re Syngenta AG MIR 162 Corn Litigation*,[14] "the parties [] asked the court to resolve a dispute concerning the format of electronic discovery to be produced" by one of the plaintiffs, which proposed "to meet the current document-production deadlines **by producing electronic discovery in native format, rather than in TIFF image format as required by the ESI Protocol Order**" [footnote omitted; emphasis added].

After already having been granted an extension to meet its production obligations, this plaintiff "produced a large number of documents in native format . . . in order to get the documents to [the defendant] as expeditiously as possible." This plaintiff claimed that converting documents to TIFF would add "substantial time to production." The defendant objected, emphasizing the requirements of the ESI Protocol Order and the inability to use per-page Bates numbering for depositions, and [this plaintiff] then asked the court "to relieve it from the production requirements of the ESI Protocol Order."

The court found this plaintiff's arguments unpersuasive and denied its request:

> First, there is no dispute that **documents in TIFF format are easier to work with and enable depositions and court proceedings to run more smoothly**. . . .
>
> Second, the ESI Protocol Order requires a party seeking to deviate from the image/TIFF-format production to "promptly" notify the requesting party as soon as it identifies a source of data to which the protocol should not apply (because it would be unduly burdensome or impractical). Here, **[this plaintiff] did not notify [the defendant] or the court before producing documents in native format**. . . .
>
> Third, [this plaintiff] has offered no evidence to support its "burdensome" and "impracticality" arguments. [emphasis added]

7.5.3 A Waived Objection

In *McDonnel Grp., LLC v. Starr Surplus Lines Ins. Co.*,[15] the defendants requested production of "all construction schedules for the Project in their native format (as native files)." Rather than objecting specifically to the requested form of production and proposing an alternative as required, the plaintiffs offered only a boilerplate objection to the overall request.

The court concluded that, "[b]y failing to object to production in native format," the objection had been waived. Moreover, the court stated that:

> . . . **the need for production in the requested, unobjected-to native format, with its associated metadata, is self-evident in this instance**. Metadata provides information such as "the author, date/time of creation and date modified." **Such information in the construction schedule context, with its frequent alterations, change orders, and time**

[14] *In re Syngenta AG MIR 162 Corn Litigation, MDL 2591,* 2018 WL 4609112 (D. Kan. Sep. 25, 2018).

> **sensitive but often disturbed deadlines, is relevant**. The PDF files chosen by plaintiff for production are merely pictures of the materials that do not provide metadata. [internal citation omitted; emphasis added]

The plaintiff also attempted to rely upon FRCP 34(b)(2)(E)(iii), which provides that "[a] party need not produce the same electronically stored information in more than one form," but the court concluded it had also "dispossessed itself of this protection" when it failed to object as required:

> To permit a responding party, in the face of a request that ESI be produced in a particular form, arbitrarily to choose some other form, **would disrupt and undermine the orderly request/response/objection/confer structure and requirements of the remainder of the Rule** concerning ESI. [emphasis added]

7.5.4 Usability and Expenses

In *Johnson v. Italian Shoemakers, Inc.*,[16] numerous issues arose regarding the plaintiffs' productions' completeness, timeliness, and format. With regard to format, the plaintiffs repeatedly produced emails in PDF format rather than in native format with metadata. The court found that to be an unjustified deviation from its discovery order:

> . . . the Court finds that Plaintiffs' August 14, 2018 production failed to comply with this Court's Order. Plaintiffs' August 14, 2018 **production consisted of emails in PDF format, which is not how emails are maintained in the regular course of business**. Further, Plaintiffs' documents were not labeled to correspond to the respective discovery request. [internal citations omitted; emphasis added]

Ultimately, the court not only ordered that production be completed as previously ordered, but also awarded sanctions:

> **The Court imposes reasonable expenses, including attorney's fees**, relating to the Motion to Compel, Motion for Sanctions, and any ongoing attorney fees related to this discovery. Further, the Court orders Plaintiffs to produce all discovery requests, including attachments, **in usable form** by the close of business on October 24, 2018. [emphasis added]

In its analysis, the court explained that the requirement in FRCP 34(b)(2)(E)(ii) to produce ESI "in a form or forms in which it is ordinarily maintained or in a reasonably usable form or forms" is satisfied "when the party provides documents that are **searchable and/or sortable by metadata fields**" [emphasis added].

7.5.5 PDFs and Metadata Requests

In *Metlife Inv'rs. USA Ins. Co. v. Lindsey*,[17] the parties' initial plan stated that "[a]ll ESI produced electronically will be produced in native format to the extent possible." Despite this, the plaintiff "generally produced documents in nonsearchable PDF format," over the defendants' repeated objections:

---

[15] *McDonnel Grp., LLC v. Starr Surplus Lines Ins. Co.*, 2018 WL 4775063 (E.D. La. Oct. 3, 2018), available at https://casetext.com/case/mcdonnel-grp-llc-v-starr-surplus-lines-ins-co.
[16] *David A. Johnson & Alda, Inc. v. Italian Shoemakers, Inc.*, 2018 WL 5266853 (W.D.N.C. Oct. 23, 2018), available at https://casetext.com/case/david-a-johnson-alda-inc-v-italian-shoemakers-inc.
[17] *Metlife Inv'rs. USA Ins. Co. v. Lindsey,* 2018 WL 5292222 (N.D. Ind. Oct. 25, 2018), available at https://casetext.com/case/metlife-investors-us-ins-co-v-lindsey-2.

> . . . MetLife concedes that the method in which it produced the documents **is not how they are kept "in the usual course of business**," as required by Rule 34(b)(2)(E)(i). Although MetLife repeatedly states that PDF is the "most usable" format, **it cites no authority showing that this satisfies its obligations under Rule 34**. Moreover, MetLife's production was **not consistent with what the parties discussed at the beginning of discovery**. [emphasis added]

The plaintiff also argued that producing the materials again in native format would impose a disproportionate burden, but the court was not persuaded:

> . . . MetLife offers no argument on that point beyond objecting to the relevance and stating that the production would be duplicative. MetLife does not discuss the volume of the additional information sought, the expense involved, or the risk of revealing any confidential or privileged information; nor has it moved for a protective order. **A request to produce documents is not disproportionate or unreasonable simply because some of the material sought has already been produced, particularly when the initial production did not conform to the rules**. [emphasis added]

Ultimately, the court ordered the plaintiff to reproduce the materials in native format and left open the possibility of an award of expenses.

7.6 PREPARING THE PRODUCTION

This preparation of a production is a collaboration between the members of the case team, the managers of any document review teams, and the internal or external technical professionals responsible for administering the chosen processing and review platforms. It typically involves four phases of activity: final pre-production checks, actual preparation, quality control, and delivery preparation.

7.6.1 Final Pre-Production Checks

The first part of this process rests with the case team, in collaboration with any review team managers. Before the actual production can be prepared, the final set of materials to be produced must be identified and final checks must be run on those materials, including:

- Checks to be sure all documents in the proposed production set are tagged as having been reviewed and as being both responsive and non-privileged (this check of tagging may be backstopped by running term searches for key privilege indicators and double-checking any results)
- Checks to make sure that the proposed production set is family group complete (if that is what has been chosen) and that all family group members have also been reviewed and determined to be non-privileged

- Checks for correct handling of email thread members and for consistent handling across near-duplicates
- Checks that all needed redactions have been correctly completed and that any protected status flags (or other indicators for endorsements) have been correctly applied
- Checks to confirm the phrasing and position to be used for required endorsements and to confirm the prefix, starting number, and position to be used for Bates numbers

Once all necessary checks have been completed, and the finalized set of materials and instructions has been confirmed, the production preparation process moves to the internal or external technical professionals responsible for administrating the processing and review platforms.

7.6.2 Actual Preparation

At this point in the process, the relevant technical professionals will engage in a series of platform-specific and production format-specific steps to actually generate the final production set for delivery, potentially including:

- Gathering together the original native files to be produced
- Generating TIFF images of them, with required endorsements
- Gathering (or creating) extracted text files for them
  - Including using OCR on redacted images to create redacted extracted text
- Programmatically renaming and organizing all natives, images, text files, etc.
- Generating load files that link all those pieces together, in the right format, with required metadata fields included and properly named
  - Including creating any custom fields and values required (e.g., protected status, request number, etc.)

Depending on the specific production format and steps required, this process can take anywhere from a few hours to a few days. In particular, generating large numbers of TIFF page images can take a significant amount of time, and for this reason, it is often begun well ahead of final production preparation to avoid last-minute time crunch.

7.6.3 Quality Control

Depending on the format choices made, the prepared production set may include thousands of native files, thousands of extracted text files, thousands of TIFF images, and a load file with numerous details about each of those thousands of files. Before delivery, this prepared production set will be subjected to some combination of quality control checks. Typically, these

are performed by the same technical professionals that prepared the production, but some may also be performed by review team managers, project managers, or the case team.

Common quality control checks for a prepared production include:

- Confirming that file counts in the prepared production match expected counts
- Spot checking a sampling of metadata fields to verify field names are right, values are right, and value formats are right
- Verifying that file path links to associated native files, page images, and extracted text files are working correctly
- Double-checking that all redactions were in fact applied to relevant page images
    - Including double-checking that extracted text for those documents has been either excluded or replaced with OCR text instead
- Verifying that endorsements have been applied to the correct documents, in the correct location, and using the correct language
    - Including verifying that Bates numbers have been applied starting at the correct number, with the correct prefix, and in the correct location

Members of the case team may also repeat some of the substantive checks performed prior to production preparation to ensure that no privileged or unreviewed materials have been inadvertently pulled into the production set during the actual preparation.

7.6.4 Delivery Preparation

Finally, once all quality control checks have been completed, the production set must be prepared for delivery to the requesting party. Options for delivery include delivery on data CDs or DVDs, delivery on flash drives or hard drives, transfer via secure file transfer protocol (SFTP), and delivery via cloud-based repositories. The primary determinant of which you use will be the size of the production:

- CDs hold around 700 megabytes
- DVDs typically hold around either 4 or 8 gigabytes
- Flash drives typically hold dozens or hundreds of gigabytes
- Hard drives typically hold hundreds of gigabytes or a few terabytes
- SFTP transfers do not have a hard limit like physical media, but are practically limited by upload and download speeds; typically, suitable for productions up to a few GB in size
- Cloud-based repositories are functionally unlimited in size

In addition to size, another factor to consider is the security of your chosen delivery method – particularly when delivering on discs or drives:

- Can you encrypt the production data you are providing on the chosen media?
- Does the chosen drive offer hardware or software level encryption?
- How will the physical media or drive be delivered?
    - By what separate method will the decryption key be provided?
- Can you protect the production data from inadvertent alteration during access?

If you are delivering via a cloud-based repository, such as a dedicated Relativity database, there are additional questions to address:

- Who, specifically, will be granted access to the repository?
- What features and abilities will be made available to them?
  - Will they be allowed to annotate documents? To export them? To print them?
- What files and formats will the database include?
  - Will it include native files? Near-native renderings? Images?
- What metadata fields will be made available in the database?
- Who will pay for hosting, for user accounts, and for any needed training?

7.7 PRIVILEGE AND PRODUCTION LOGS

As you approach the end of your production efforts, there are two additional steps that should be taken prior to delivery of the prepared production. First, if any materials have been withheld due to privilege or work product protection, those materials will need to be documented in a privilege log. Second, for your own records, you should prepare a production log documenting your production.

7.7.1 Privilege Logs

Protecting privileged materials from inadvertent disclosure is of paramount importance during discovery, both because attorneys have an ethical duty to protect client confidentiality (see, e.g., ABA Model Rule of Professional Conduct 1.6[18]) and because inadvertent disclosures can lead to privilege waiver if reasonable steps to prevent the disclosure weren't taken (see Federal Rule of Evidence 502(b)[19]). The final step in that privilege protection process is the preparation of some type of privilege log to accompany your production set delivery.

FRCP 26(b)(5)(A)[20] provides the basis for this requirement in federal courts:

> (A) Information Withheld. When a party withholds information otherwise discoverable by claiming that the information is privileged or subject to protection as trial-preparation material, the party must:
>
> (i) expressly make the claim; and
>
> (ii) describe the nature of the documents, communications, or tangible things not produced or disclosed – and do so in a manner that, without revealing information itself privileged or protected, will enable other parties to assess the claim.

The preparation of traditional privilege logs can be a time consuming process, since each individual document withheld must be recorded, the claim for it articulated, and an adequate description written, and since, in larger cases, the total number of documents requiring logging

[18]ABA Model Rules of Prof'l Conduct R. 1.6 (2021), available at https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_6_confidentiality_of_information/.
[19]Fed. R. Evid. 502(b), available at https://www.law.cornell.edu/rules/fre/rule_502.

can be quite high. As a result, it is common to begin planning and preparation for privilege log creation during the document review phase of an eDiscovery project.

It is common during document review to have reviewers designate not just general privileged status, but also to select the applicable legal basis from pre-written options so that those fields can be automatically populated during privilege log creation. To facilitate this, many document review platforms include features for privilege log creation, including the ability to automatically populate the log with key details about documents (e.g., date, file name, file type, sender, recipients, subject line, etc.).

There are also alternative approaches[21] to privilege log preparation that focus on defining categories of materials withheld rather than creating document-by-document log entries. In the face of ever-increasing data volumes, these categorical approaches are growing in popularity and finding favor in some courts.[22]

7.7.2 Production Logs

In addition to creating the privilege log that you will provide with your production set delivery, it is also important to create and maintain a production log for yourself. In this era of large ESI volumes, it is common to complete multiple productions on a rolling basis, providing responses to different requests or materials from different sources as work on them is completed. Additionally, matters with an open-ended period of relevance may require supplemental productions to be made as new responsive materials are generated.

A production history log documents all of the details you might need to know later (or be able to demonstrate to someone else later) about all of those productions. Key details to document include: what you produced, what formats you produced it in, when you produced it, how you delivered it, to whom who you delivered it, the requests to which it responded, and the Bates ranges it contained.

[21]Hon. John M. Facciola and Jonathan M. Redgrave, *Asserting and Challenging Privilege Claims in Modern Litigation: The Facciola-Redgrave Framework*, 2009 FED. CTS. L. REV. 4 (Nov. 2009), available at https://www.fclr.org/fclr/articles/html/2009/facciolaredgrave.pdf.
[22]*Several Courts Allow Categorical Privilege Logs* , MCGUIREWOODS, https://www.mcguirewoods.com/client-resources/privilege-ethics/Privilege-Points/2021/1/several-courts-allow-categorical-privilege-logs (Jan. 20, 2021).

7.8 KEY TAKEAWAYS

There are six key takeaways from this chapter to remember:

1. Effective production of ESI is both a requirement of the rules and one element of fulfilling an attorney's duty of technology competence for eDiscovery.
2. Productions can be made in paper, near-paper, native, and near native formats (or combinations thereof), and they may require decisions about load files, metadata, unitization, redactions, endorsements, scanned physical documents, and other specifics.
3. The production format and related specifics should be negotiated between the parties as part of their initial meet-and-confer, but when no agreement has been negotiated, parties can also later request responses in a particular format, object to a requested format, and if necessary, seek orders to compel or protect.
4. In the absence of an agreement, request, or order otherwise, ESI must be produced either:
    a. Formatted as it is kept in the ordinary course of business (*i.e.*, native format), organized as it is kept in the ordinary course of business.
    b. In another reasonably usable format (*i.e.*, one that is searchable and that is sortable by metadata), labeled to correspond to the categories in the request.
5. Thorough quality control checks should be performed to ensure: that the right materials are designated for inclusion in the production, that those materials (and only those) actually appear in the prepared production, and that the prepared production matches the required production specifications.
6. In addition to the production set deliverable itself, you must also prepare a detailed privilege log for the requesting party and a detailed production log for yourself.

Unit 2

Intermediate eDiscovery

Chapter 8 - Measure Twice, Discover Once: eDiscovery Project Scoping and Planning

Managing eDiscovery matters without a plan is like trying to navigate a major city without directions – chaotic, inefficient, and unlikely to get you where you wanted to go. Taking the time to scope and plan will help you understand where you need to go and how best to get there.

Chapter 9 - Hold On: Get a Grip on Conducting Effective Legal Holds

As eDiscovery expands to include social media, mobile apps, collaboration tools, and more, understanding and conducting effective legal holds is more critical than ever. Done properly, legal holds set the stage for efficient and accurate matter management. Conducted without direction, legal holds can become a minefield for litigants.

Chapter 10 - Beyond the Four Corners: Evolving Electronic Documents

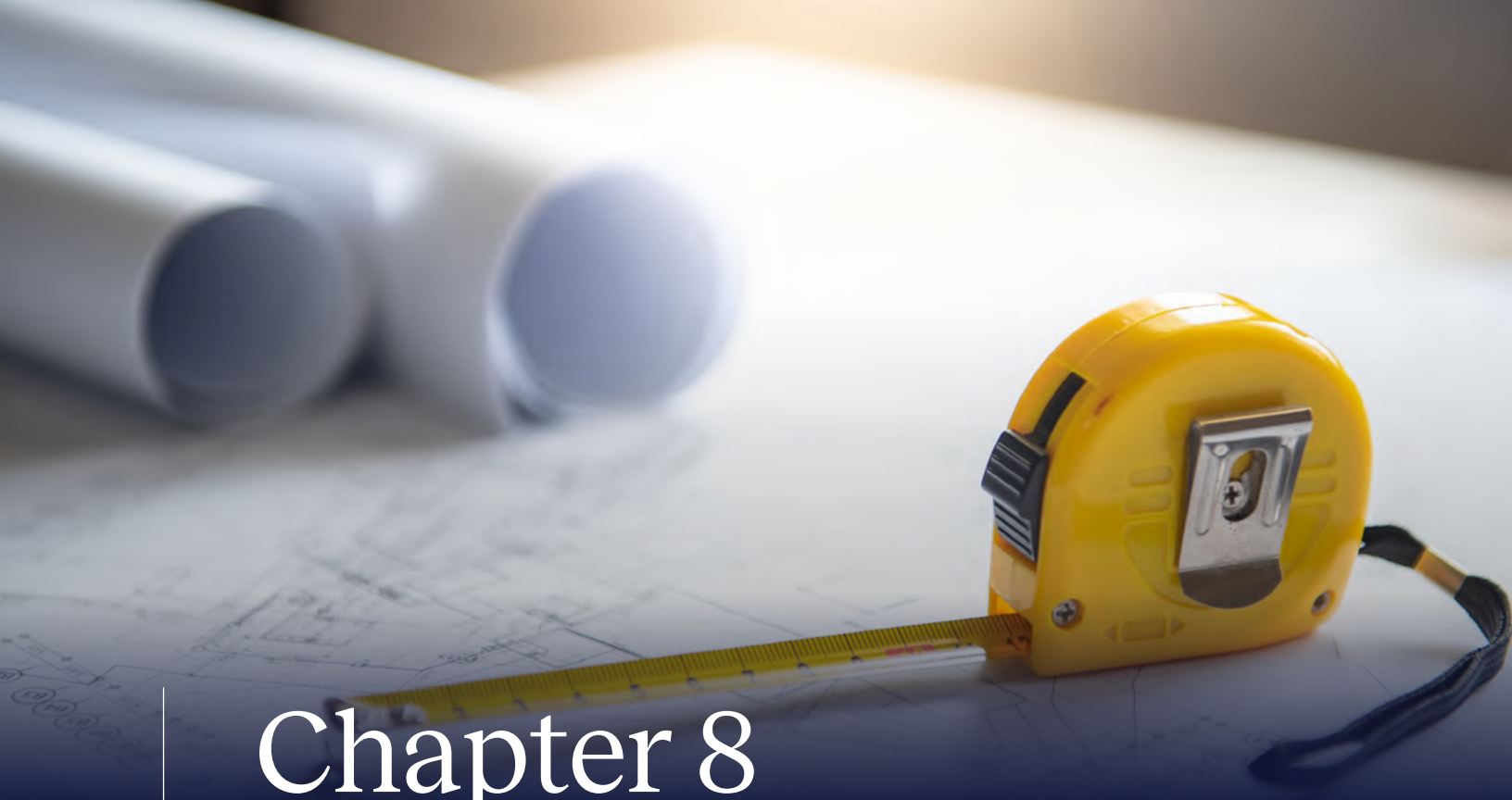
The definition and boundaries of "document" are changing, as new source and file types proliferate and custodian behavior changes. New challenges abound, like modern attachments, dynamic content, endless threads, and more. This chapter reviews these new technical challenges, the legal ambiguities they create, and the ways practitioners are approaching them.

Chapter 11 - Sampling Techniques for Litigation and Investigations

Despite years of discussion in the eDiscovery industry about the power and importance of sampling techniques – particularly in the context of technology-assisted review (TAR), many practitioners remain unfamiliar with what they can accomplish with them, and when, outside of TAR, they might do so.

Chapter 12 - An Embarrassment of Riches: Analytic Tools and Techniques

The tide of data never stops rising, and the sources of data never stop multiplying, and legal practitioners must somehow find a way to analyze it. Finding a way that is efficient and effective requires understanding the tools and techniques available to you so you can leverage the right ones.



Chapter 8

Measure Twice, Discover Once: eDiscovery Project Scoping and Planning

About this Chapter

In this chapter, we will discuss various aspects of effective eDiscovery project planning to equip you with the knowledge you need to reduce your chaos, your costs, and your risks. We will review: initial scoping activities; investigation activities; volume and cost estimation; and roles and communication.

8.1 CLICHÉS, CHAOS, AND EDISCOVERY PROJECT PLANNING

We have a lot of maxims in English about the value of preparation, like "measure twice, cut once" or "a stitch in time saves nine." Perhaps the most well-known aphorism of this type is "an ounce of prevention is worth a pound of cure," which dates back to an anonymous letter on the importance of fire safety that Benjamin Franklin published[1] in the Pennsylvania Gazette in the early 18th century. Though Franklin was writing of fire safety and the consequences of laxity in that area, he might as well have been writing about eDiscovery project planning and its risks. As any experienced eDiscovery practitioner will confirm, being forced to risk your neck jumping out the window of a burning house, because of something tiny you missed while hurrying around, can be a pretty good analogy for the eDiscovery experience.

8.1.1 eDiscovery on Fire

eDiscovery is undeniably challenging. Data volumes continue to multiply, data types continue to diversify, and data custodians continue to modify their tools and practices. Couple this daunting set of variables with an ever-expanding set of eDiscovery tools and services available to be leveraged, add time pressure and an adversarial process, and you have a perfect recipe for chaos, uncertainty, and small (but important) things getting missed.

As Franklin's letter and the other maxims tell us, the reliable way to reduce the risk of such errors is to take the time for proper planning before rushing headlong to action. To be sure, planning an eDiscovery project is an iterative process that overlaps and intersects with other early project activities, but investing the time and effort required for effective planning, from the beginning (and throughout those early phases), will produce downstream benefits, including saved time, saved money, reduced risk, and increased defensibility. The clichés became clichés for a reason.

8.1.2 Evolving Expectations

State bars and industry organizations formally recognize the importance[2] of competent preparation and planning to meet the technical and logistical challenges of our current eDiscovery reality. For example, California's Formal Opinion on attorneys' duty of eDiscovery competence[3] specifically articulates that attorneys (or attorneys working with the assistance of qualified experts) need to be able to:

- "Initially Assess E-Discovery Needs and Issues, If Any"
- "Analyze and Understand a Client's ESI Systems and Storage"
- "Advise the Client on Available Options for Collection and Preservation of ESI"
- "Identify Custodians of Potentially Relevant ESI"
- "Engage in Competent and Meaningful Meet and Confer with Opposing Counsel Concerning an E-Discovery Plan"

Each of these requirements is part and parcel of effective eDiscovery project planning, making the ability to do such planning a formal requirement in California. The EDRM organization, too, in

[1]Benjamin Franklin, *On Protection of Towns from Fire*, The Pennsylvania Gazette (Feb. 4, 1735), available at https://founders.archives.gov/documents/Franklin/01-02-02-0002.
[2]Robert J. Ambrogi, *Tech Competence*, LAWSITES, https://www.lawnext.com/tech-competence (2021).
[3]The State Bar of California Standing Committee On Professional Responsibility and Conduct, *Formal Opinion No. 2015-193* (June 30, 2015), available at https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL 2015-193 %5B11-0004%5D (06-30-15) - FINAL.pdf.

their EDRM Project Management Framework[4] ("EPMF") emphasizes the critical importance of the early planning phases to overall project success:

> **The Scoping Phase lays the foundation for project success by aligning all teams to a common vision and goals**. By defining the project scope and providing an overview of the context and constraints, this phase lays the foundation for project planning in the following phases. [emphasis added]

Their EPMF scoping phase is then followed by a preliminary planning phase and a detailed planning phase.

Checklists

Throughout this chapter, we will also review checklists of key items for each phase in the process. Checklists may sound like simple things, but when properly implemented, their value is real. For example, doctors starting to follow checklists for key tasks in one program saved 1,500 lives and $175 million in just the first eighteen months.[5] If checklists are a match for modern medicine, they're more than a match for eDiscovery.

8.2 INITIAL SCOPING ACTIVITIES

We don't spend a lot of time talking about imagination in legal practice, but it's pretty essential to effective project scoping. Unless you're working in an eDiscovery operation with Level 5 maturity[6] (i.e., with extensive aggregated data about past projects, matter types, etc.), each new project is going to be fairly opaque to you at the outset. You will know the general legal subject matter, the essential event(s) giving rise to the issue, and any individually named defendants within the organization. Beyond that, however, anything and everything (or nothing) might be relevant.

The first step, then, must be brainstorming to figure out what and who might be relevant. Doing this effectively requires collaborating with: individuals with direct knowledge of the relevant legal issues; individuals with direct knowledge of the relevant factual issues; and, individuals with direct knowledge of the organization's individual and enterprise IT resources. Essentially, in house counsel, outside counsel, internal IT, any external collection resources, and any relevant senior employees all need to be looped into this exercise in imagining what might exist.

Starting with what you know about the type of matter, the underlying facts, and the key players (typically based on a complaint or preservation notice), you and your collaborators must extrapolate what types of relevant materials are likely to exist within the organization and where (or in whose custody) they are likely to be. This should include consideration of, both the information and materials you will want to see and use, and the information and materials you anticipate the opposing party will request. This is the general flow of inquiry:

[4] *Project Management Guide*, EDRM, http://www.edrm.net/frameworks-and-standards/edrm-model/projectmanagement/(2020).
[5] Atul Gawande, *The Checklist: If something so simple can transform intensive care, what else can it do?*, The New Yorker (Dec. 2, 2007), available at http://www.newyorker.com/magazine/2007/12/10/the-checklist.
[6] *Capability Maturity Model*, Wikipedia, https://en.wikipedia.org/wiki/Capability_Maturity_Model#Levels (Mar. 18, 2021).

- What events are in dispute or under investigation?
    - What questions do you have about those events?
        - What materials might help you answer them?
    - What questions are parties-opponent likely to have about those events?
        - What materials might they imagine exist and request?
- What are the elements of the legal claims and defenses in the matter?
    - What types of documentary evidence might help establish or refute them?
        - Where or in whose custody might such evidence be?

This process can be aided by checklists of potential sources, like those used for custodian surveys and interviews, but it is a fundamentally imaginative exercise. Imagine the events at issue in the context of normal organizational operations and think about what might have been generated:

- Might there be departmental records, like HR files?
- Could there be useful data about the events in your ERP systems?
- Would employees have discussed the events via the internal chat client?
- Maybe the relevant office used shared network folders?
- Perhaps copies of deleted records exist on back-up tapes?

Additionally, it is beneficial to imagine what distinctive characteristics relevant materials from these sources might bear, i.e. how you would try to find them if searching for them in a collected data set:

- Are you seeking evidence of intent in communications between certain employees?
- Are you looking for evidence of internal awareness in executive meeting minutes?
- Will relevant documents contain certain keywords, like a name or project code?
- Are you looking for contracts executed with a particular party or on certain dates?
- Are you looking for metadata evidence of an employee altering key documents?

Having some ideas about distinctive characteristics of this type will be helpful to you as you move on to the investigation activities we will discuss below.

8.2.1 Prioritization

Once you have finished your brainstorming exercise and have a list of potentially-extant relevant materials, likely places to look for them, and distinctive characteristics you might use to identify them, your next step is to prioritize these potential materials, sources, and custodians to guide

your subsequent activities and allocation of resources. Prioritization of the items on your list – whether they are IT systems, departmental systems, or individual custodians' devices – should be done based on three key criteria:

1. How likely it is that the source actually contains relevant materials
2. How important and useful those materials would be, if they do exist
3. And, how likely those materials are to be requested by parties-opponent

Obviously, the greater the likelihood it exists, the greater its potential utility, or the greater the likelihood it will be requested, then the greater its priority should be (with items that score high on all three metrics at the top of the list). This prioritized list of potential sources can then be broken into three tiers that can be used to guide prioritization of subsequent steps, determination of proportional levels of effort, or phases for a phased discovery plan:

- Tier 1 – Materials that must be sought (key sources/custodians)
- Tier 2 – Materials that may need to be sought (secondary sources/custodians)
- Tier 3 – Materials that may not need to be sought (tertiary sources/custodians)

Once you have completed this prioritization and grouping, you are ready to begin the investigation activities we will review next.

Initial Scoping Checklist

1. What are the events at issue, and what are the legal claims?
2. What internal and external individuals should be involved in brainstorming?
3. What questions do we have about the underlying events? What questions will they?
4. What materials might answer those questions for us and for them?
5. What are the elements of the relevant legal claims and defenses?
6. What materials might establish or refute each of them?
7. What individuals, devices, departments, or systems might contain them?
8. What distinctive characteristics might help you search or filter for them?
9. What is the relative priority of each potential source identified?
10. What grouping into tiers should be applied for planning subsequent steps?

8.3 INVESTIGATION ACTIVITIES

Once you have collaborated with knowledgeable individuals to brainstorm hypothetical materials and potential sources (and prioritize them), you are ready to begin investigating the facts on the ground to bridge the gap from your imagination to actual reality. A variety of investigative options

are available for accomplishing this, including: targeted interviewing, data mapping, surveying, and sampling. Which one (or more than one) will be most useful to you will depend on your circumstances – in particular, your expected number and types of sources. For example:

- The larger your project, the more investigative steps you'll need to take
- The more systems and sources by count, the more useful a data map is
- The more custodians by count, the more useful surveys and samples are

8.3.1 Targeted Interviews

Targeted interviews are the easiest investigative step and a common first one. In this context, conducting targeted interviews is like conducting a limited number of custodian interviews with key personnel. This process is typically less formal (i.e., no full script) and less complete (i.e., most individual custodians aren't included) than the official custodian interview process, which will come later in the project. (As we noted above, planning an eDiscovery project is an iterative process that overlaps and intersects with other early project activities.)

Your goal in the targeted interviews is to review your list of prioritized, hypothetical materials with individuals that have knowledge of the potentially relevant enterprise, departmental, and third-party systems – as well as the computers and devices typically issued to individual custodians within the organization – to confirm or deny your assumptions and gather the information you will need to scope and plan further.

Targeted Interviews Checklist

1. Who is familiar with enterprise, departmental, and/or third-party systems?
2. Who can provide details about employee computers, devices, and usage?
3. Who else might be able to confirm or deny assumptions about what's there?
4. What details would you ideally like to know about the potentially relevant sources?
5. Has everything from your prioritized list been reviewed, detailed, and documented?

8.3.2 Surveys

As we noted above, the investigative options you need to undertake to test your project assumptions will depend on the specifics of your project, especially on its scale. Larger or more complex projects will require more – and more ambitious – investigative efforts. Surveys are particularly useful and important in projects that feature a large number of potential custodians.

Surveying in this context, like targeted interviews, is part and parcel of what would normally be your full, pre-collection custodian interview process. And, like targeted interviews, it is worth thinking about surveying as more than just collection planning. Beyond just documenting

Surveying Checklist

1. Do we need additional information from many potential custodians?
2. Would self-reported answers be sufficient for our current purposes?
3. Given the questions and recipients, what format makes the most sense?
4. How will answers be collected, aggregated, and made useful for planning?
5. How will completion tracking, reminder issuance, and follow-ups be done?

8.3.3 Data Mapping

Your next investigative option is data mapping. Data mapping is the process of "mapping" the various data stores and sources in an organization. Many organizations do some version of this already for non- legal purposes. For example, the IT or IS department may have "maps" of the organization's servers, computers, and enterprise systems, along with directories of installed software. A data map for the legal activities like eDiscovery, however, is a related but distinct thing. This kind of data map needs to combine system details, content details, and other key details (e.g., who owns it, any built-in export tools, etc.) to facilitate preservation and collection.

Ideally, data mapping for legal activities would be undertaken on a proactive, organization-wide basis rather than in response to a specific matter, but engaging in some targeted, reactive data mapping is better than none and well worth doing. (And, it can form the basis for proactively proceeding to organization-wide data mapping once the current matter is has concluded.)

In this context, you would be working your way down your potential materials/hypothetical sources list, reviewing them with relevant individuals (from IT/IS, Records Management, etc.) and reviewing relevant (information systems and records management) documentation, attempting to flesh out that list with concrete details. What you will be attempting to build is less a literal map than a spreadsheet or matrix. Your final product will be a searchable, sortable, filterable reference tool listing sources in rows and relevant details about them in columns.

Things you may need to know about each source include:

- Source type (e.g., enterprise, department, individual custodian)
- Owner/manager of source (e.g., specific IT contact, department manager, or custodian)
- Types, models, and years for source's hard- ware systems or custodian devices
- Versions, years, and other details for source's relevant software

- Available native search and export tools/features, if any, and relevant details
- Limitations of such tools, if any (e.g., one mail- box at a time, can't search nested content)
- Desired materials expected to be there (including expected formats, dates, etc.)
- Expected volume of materials from source (e.g., record count, file volume)
- Relative priority (and sensitivity, if applicable) of those desired materials

Gathering and organizing this information (or as much of it as time and circumstances permit) will enable you to scope and plan your needed preservation and collection activities with a high degree of precision. You will know which sources can be handled internally and which require specialists, which are likely to present technical challenges and which can be had quickly, and which are most likely to be really important and which are most likely to be duplicative.

Data Mapping Checklist

1. What kinds of hypothetical systems and devices are implicated by your list?
2. What details about those systems would it be most useful for you to know?
3. From where and whom within the organization could those details be gathered?
4. Have all relevant (and reasonably obtainable) source details been gathered?
5. Are those details consistently described in a manipulable spreadsheet or matrix?

8.3.4 Sampling

The final investigative option we'll review is sampling. In the land of eDiscovery, sampling is used to refer to both judgmental and statistical sampling. In this early project planning phase, both kinds of sampling can be useful.

Judgmental sampling is the informal process of looking at parts of something large to get an anecdotal sense of the whole. For example, attorneys are engaged in judgmental sampling when they run a variety of instinctively-selected search terms in a document collection to familiarize themselves with what's there. Judgmental sampling is also what you're doing when you select key individuals for targeted interviews, using them as proxies for the whole list of hypothetical custodians.

More importantly, though, judgmental sampling is a way to learn about what's on sources and systems that, unlike custodians, cannot self-report to you. This kind of judgmental sampling might take a variety of forms, such as:

- Testing/searching electronic mailboxes to test relevance before collection
- Indexing/evaluating some backup tapes to test for unique materials in backups
- Collecting representative custodians' lap- tops (or phones, etc.) to test relevance

Statistical sampling is the more formal process of taking simple random samples of sufficient size to reliably estimate properties of the whole set. For example, reviewing and coding 2,400

randomly selected documents from a million-document set to estimate, with a confidence level of 95% and a confidence interval of +/-2%, how much of the total is relevant (or privileged, confidential, etc.).

Depending on your project's scale and timeline, you may proceed from judgmental sampling to statistical sampling, by loading and coding formal samples from your initial, test collections. If you need to say with certainty whether a category of devices or sources is worth pursuing further, formal sampling of one or more devices' contents can provide that certainty, and if those contents are voluminous, formal sampling will do so far faster than broad review.

These sampling techniques are especially important in this era of increased focus on proportionality. Negotiations with opposing parties often happen in parallel with internal project planning and other early activities, and negotiations about the appropriate scope and scale of discovery are always more effective when assumptions can be backed up (or disproved) with actual facts and examples. This provides an additional use for, and benefit from, your investigative efforts. For years, judges have been emphasizing to parties the importance of using sampling to flesh out facts about what is and isn't actually there instead of fighting over theories about what might be.[7]

Sampling Checklist

1. Do we need additional information about what's on various sources or systems?
2. Do we need support for discovery negotiations as well as our own project planning?
3. Would judgmental sampling of one or more of the devices be sufficient?
4. Is formal statistical sampling needed to take more specific measurements?
5. How will the sampling process, including decisions and rationales, be documented?

8.4 VOLUME AND COST ESTIMATION

Once you have completed your initial planning and completed your investigation activities to validate and flesh out your initial assumptions, you should be equipped with enough information to proceed to estimations of project volumes and potential costs. At a minimum, you need a reasonably accurate count of:

- Custodians requiring collection
- Devices per custodian requiring collection
- Mailboxes and network shares requiring collection
- Enterprise or departmental systems requiring collection
- Cloud-based sources requiring collection (e.g., Slack, Teams)
- Backup tapes or other loose media requiring collection

[7] *See, e.g., Pippins vs. KPMG LLP*, 279 F.R.D. 245 (S.D.N.Y. Feb. 2, 2012), available at http://pdfserver.amlaw.com/legaltechnology/Pippins_v_KPMG_Order_20120203.pdf.

8.4.1 Volume Estimation

At this point, you should have some sense of how large each category of sources is (i.e., laptop size, mailbox size, etc.) and how broadly you expect to have to collect (i.e., full images vs. logical images vs. pre- filtered collections/exports). With this information, making an educated guess as to your initial collected volume becomes straightforward:

Custodians x (Sum of Issued Devices' Typical Sizes)
+ Mailboxes x Typical Size
+ Network Shares x Typical Size
+ Sum of Enterprise and Departmental Systems' Sizes
+ Estimate of cloud-based source volumes
+ Backup Tapes/Storage Media x Tape/Media Sizes = Approximate Total ESI Volume to Collect

Once you have this number, you will need to make some additional assumptions and adjustments to project your likely downstream volumes.

First, you'll need to consider the expansion of the collected data volume that will occur at the beginning of processing. For example, your collected data volume will include some number of compressed container files (e.g., ZIP, RAR, etc.), each of which will expand into one or more files of larger size than when compressed. Other types of compressed and nested content also exist (e.g., local PST and OST email stores), and during processing all will be fully expanded so each element can be individually normalized, tracked, and reviewed. In particular, cloud-based collaboration tools are prone to significant post-collection expansion. The amount of expansion can vary widely – from as little 10%, to more than 40%, up to 1,000% in some cases – depending on just what was in the original collection. Collections from collaboration tools, in particular, tend to expand dramatically.

Second, you'll need to consider the immediate reductions that will occur from de-NISTing, deduplication, and the application of any objective filters:

- **De-NISTing**: It is standard practice to de- NIST each collection to eliminate system, software, and utility files that can have no bearing on the matter at hand. Just how much material will be eliminated depends on how narrowly or broadly the collection was done. Full disc images will be greatly reduced in volume; targeted collections of user files will not.
- **Deduplication**: It is also standard practice to globally deduplicate each collection so that only one copy of each record need be reviewed, managed, etc. Modern discovery software makes the tracking of each place a duplicate was, as well as their later restoration and handling, simple matters. Email heavy collections will see the most volume reduction, as every party to an internal email communication may have duplicate copies of every message between them.
- **Objective Filtering**: It is also very common for a new collection to have some objective filtering applied during processing based on the scope of the case or the negotiated limits of discovery – for example, the application of date restrictions or file type restrictions. How much additional impact these filters will have on volume will depend on how narrowly targeted the initial collection process was.

As both parties and collection tools have grown more sophisticated in recent years, the trend has definitely been towards smaller, more-targeted initial collections that therefore reduce less during this phase. Additionally, it should be noted (when estimating volume for hosting costs) that the final, post-processing volume will expand slightly again when loaded into a review platform to accommodate the review platform's database file, extracted text files, etc.

A variety of tools and analyses are available to help you select your assumptions and do these sorts of estimations. The EDRM organization has collected several free calculators here,[8] and their own EDRM Data Calculator[9] is a good place to start. For moving on to cost estimations and other downstream planning, you may need to estimate not only total volumes, but also potential document or page counts (for review cost estimation). The number of documents or pages in a given gigabyte of collected ESI can vary dramatically depending on the source type and the collection method. Consult with your collection and processing service providers to determine a reasonable estimate for your specific circumstances.

Volume Estimation Checklist

1. How many of each kind of source do we have?
2. How large do we expect each individual source to be?
3. How broadly or narrowly do plan to collect for this matter?
4. How much volume expansion do we anticipate during processing?
5. How much reduction from de-NISTing, deduplication, and objective filters?

8.4.2 Cost Estimation

At this point in your process, you have completed your initial planning, completed your investigation activities to validate and flesh out your initial assumptions, and you've completed your project volume estimations. You now have enough information to also do cost estimation:

- Source types and counts for estimating collection costs
- Projected collected volume for estimating processing costs

[8] *Budget Calculators*, EDRM, http://www.edrm.net/resources/budget-calculators/ (2020).
[9] *EDRM Data Calculator*, EDRM, http://www.edrm.net/resources/budget-calculators/edrm-data-calculator/ (2020).

- Projected post-processing volume for estimating hosting costs
- Projected file/document counts for estimating review costs

As with volume estimation, cost estimation is now a straightforward process of multiplying your projected volumes and counts by your preferred service provider's price prices. Several of the calculators linked above for data volume estimation can be used to help you estimate pricing as well, and many service providers also offer their own calculator built to reflect their specific pricing model.

Estimation of the review costs portion does require some additional work. Simply dividing your projected document count by 50 documents per hour to get a total number of hours of review to be performed will not give you an accurate estimate. Instead, you must consider a number of additional variables:

- How much do you believe the collection can be further reduced during ECA?
  - *By using searching, sampling, filtering, clustering, etc.*
- Will you use near-duplicate identification and email threading to reduce further?
  - *They both reduce volume and increase review speed*
- Will review be traditional or technology-assisted (i.e., TAR or CAL)?
  - *The former takes more first-level hours, the latter more QC*
- How much training, oversight, and quality control time will be needed?
  - *Management, oversight, and QC in- crease exponentially with project size*
- How much privileged material do we anticipate needing to code and log?
  - *Assume a minimum of 5-10% privileged materials requiring logging*
- Do we expect many spreadsheets, technical drawings, or other difficult documents?
  - *If there are enough, establishing specialized workflows can save time*

All of these variables will affect how much must be reviewed, how fast it can be reviewed, and how many labor hours the total effort will take. An experienced eDiscovery project manager can help you think through these options and their effects.

Cost Estimation Checklist

1. What are our projected volumes before and after processing?
2. How much additional volume reduction do we expect after processing?
3. What review options and methods do we expect to employ in this matter?
4. How much privileged or technically-challenging material do we expect?
5. What price list, bundle, or model are we using for this matter?

8.5 ROLES AND COMMUNICATION

As you transition from your planning and estimation activities into the full eDiscovery project, taking time to predefine key roles and communication guidelines can save significant time and confusion later. Even a modestly-sized eDiscovery project is likely to involve individuals from:

- The client organization:
  - In-house counsel and support staff
  - IT and records management personnel
- One or more law firms:
  - Case team attorneys and support staff
  - Litigation support personnel
  - Internal or contracted review team personnel
- One or more eDiscovery services providers:
  - Forensic collection personnel
  - Data processing personnel
  - Document review personnel
  - Project management and support personnel

That is a lot of groups and individuals – each with distinct perspectives and priorities – to keep coordinated and moving towards the same goals. The clearer the roles and guidelines established at the beginning, the easier that will be to do.

8.5.1 Primary Points of Contact

eDiscovery projects generate phenomenal amounts of intra- and inter-organizational communications, especially during the first few phases of activity. To keep that communication flowing smoothly, it is useful to identify a single, primary point of contact for each organization. The majority of communication with that organization about the project should go through this individual, and they should be copied on any communications going directly to others on their team. This individual is typically a project manager for the service provider, a paralegal or litigation manager for the client organization, and a junior attorney or paralegal for the law firm. The identified individuals function as air-traffic controllers, ensuring that all traffic gets directed to the correct people within their respective organizations. They also serve as early warning systems that can keep an eye out for potential issues requiring priority or escalation.

8.5.2 Delegations of Key Authority

Another extremely common challenge of early discovery phases is getting key decisions made in a timely fashion. For example:

- Device acquisition decisions during on- site collection efforts

- Exception handling decisions during data processing
- Batch coding decisions during early case assessment
- Tagging palette change decisions during early review

Delays in any of these decisions – or many others – can cost money, as people and resources sit idle awaiting instructions. These delays can be mitigated or avoided by predetermining the scope of authority being delegated to key team members at each organization. Can the junior associate make these decisions without approval from the partner? Can the law firm without approval from the client organization? At a minimum, a designated decision- maker for on-the-fly collection scope changes should be identified before full-scale collection is begun.

8.5.3 Issue Escalation Paths

In any eDiscovery project, unexpected issues are inevitable: a custodian will fail to cooperate, a server will go down, a last-minute scope change will be made, etc. When those issues arise, it will be necessary to escalate the issues beyond the primary contact people for each organization and past the first-level decision-makers handling day- to-day activities. Senior management of a service provider may need to step in to ensure resolution, senior partners may need to make difficult decisions, or the AGC or GC may need to get involved to approve additional expenditures.

Knowing in advance how these sorts of issues should be escalated, and to whom, can save time, money, and frustration when those issues arise. You will want to know:

- Who are the after-hours points of contact at each organization and what are the preferred methods of contacting them?
- What is the escalation path at each organization, when primary points of contact or afterhours points of contact cannot be reached or cannot provide resolution?
- Who has final project authority and responsibility at each organization?

8.5.4 E-mail Communication Rules

Because of the large volume of email communication that will go on, and because of the legal significance of much of that communication, it is also important to establish some rules for that communication:

- First, how should the emails be labeled?
  - This includes, both any required privilege and confidentiality warnings that need to be applied to satisfy legal requirements, and any standardized subject line flag (e.g., matter name and number) to aid later identification, organization, and searching.
- Second, how should the emails be stored?
  - Since all of the emails are now relevant to the matter (even if all protected as privileged or work product), they must be preserved. Retention expectations and storage instructions (e.g., Outlook foldering and folder labelling instructions) should be communicated.
- Third, are there any restrictions on who can be included on project emails? Or on what can be discussed in them?
  - For example, some law firms or service providers may have a firewall between

different internal project teams to avoid potential conflicts, or an organization may deem its proprietary business information too sensitive for any discussion in email.

8.5.5 Other Documentation Rules

Finally, you will want to establish some rules for any project documentation beyond email communications. These rules should specify how non-email materials should be labeled, stored, and screened, just as you have for email communications. Beyond that, these rules should also cover any documentation that needs to be generated. For example:

- Are there recurring reports (e.g., weekly or monthly progress against budget) that need to be generated and distributed?
- Are confirmation emails to be sent documenting project decisions? Change logs?
- How should decision-making about culling and coding during ECA be recorded?

Establishing these rules from the outset will ensure that you have the materials and information you may need later to explain or defend the conduct of the project and its key decisions.

Role and Communication Checklist

1. Who will function as each organization's primary point of contact?
2. Who at each organization will have authority to make decisions?
3. How and from whom can after-hours assistance be obtained?
4. What is the escalation path for each organization?
5. Who is the ultimate decision-maker for each organization?
6. What protective language or identifiers should be used in email?
7. Are there any limits on the recipients or topics for email?
8. What are the storage and retention rules for email?
9. What are the labeling, storage, and retention rules for non-email documentation?
10. What recurring documentation, if any, needs to be generated?

8.6 A FINAL RECOMMENDATION

One final piece of advice for effective eDiscovery project scoping and planning: seek help from experienced practitioners early and often. It is quite common for organizations not to involve an eDiscovery service provider or independent expert in their eDiscovery project efforts until much of the initial scoping and planning has already been done – sometimes after the meet-and-confer has already occurred and an agreement has already been negotiated. Unfortunately, at that point it's already too late to avoid some of the pitfalls discussed above, and the negotiated agreement may not even be technically feasible.

So, it's worth remembering: you can always opt to involve a service provider or expert practitioner for a few hours of early consultation and planning assistance to help you check your blind spots and get off to a strong start, without committing yourself to outsourcing every phase of the project.

8.7 KEY TAKEAWAYS

There are six key takeaways from this chapter to remember:

1. **eDiscovery planning reduces risk and cost**. The most expensive mistakes are the ones made at the beginning, which can result in lost evidence or large-scale do-overs. Spending some time engaged in effective scoping and planning is an investment in avoiding those issues. Implementing some standard checklists to ensure the completeness and consistency of the scoping and planning process from matter to matter can be very effective.
2. **Imagine events in context, what might exist, and what's needed.** Thinking thoroughly through what is likely to exist, what you are likely to need/want, and what your opponent is likely to need/want is the best way to avoid key materials being missed or lost. Once you have a list, prioritize it by likelihood it exists, importance if it does, and chances of it being requested by the other side.
3. **Investigate as thoroughly as you need to test your assumptions.** Leverage one or more of targeted interviewing, surveying, data mapping, and sampling to confirm what exists and gather useful details. Data maps are most helpful for wrangling enterprise and departmental systems; surveys are most useful for wrangling large numbers of potential custodians; and sampling is a powerful way to replace hypotheticals with evidence and examples for planning and negotiation.
4. **Use your gathered data to estimate volumes and document counts.** Remember that volume will expand during processing – sometimes dramatically – and then reduce (some) due to objective filtering. Remember that loaded, hosted volume will increase again (slightly). Estimations of volumes and document counts can be used to estimate the costs for each project phase, but more variables must be considered when estimating review costs.
5. **Define roles and communication as clearly as possible.** eDiscovery projects typically involve numerous individuals from numerous organizations. Keeping everyone coordinated and moving towards a common goal is easiest when there are designated primary points of contact for each organization, predetermined delegations of authority, clear escalation paths, and guidelines for communication and documentation in place.

Chapter 9

Hold On:
Get a Grip on Conducting Effective Legal Holds

About this Chapter

In this chapter, we will provide you with information to help you meet those challenges and achieve effective, consistent, and well-documented legal holds for your organization. We will review the duty of preservation, the essential elements of an effective hold, the processes and policies you should consider, the tools for hold issuance and tracking, and other issues.

9.1 A LEGAL HOLD IS JUST A LETTER, WHAT COULD GO WRONG?

Legal holds remain a common source of issues for litigants, particularly with regard to the spoliation that can follow an ineffective or nonexistent legal hold and the question of whether reasonable efforts to preserve were taken. Some examples of the potential consequences include:

- **Obstruction of justice charges**
  - United States v. Volkswagen AG, No. 16-CR-20394 (USDC EDMI Jan. 11, 2017) (Third Superseding Information ¶¶ 22-25 and Plea Agreement at ¶¶ 73-82)[1]
- **$2.7 million award of fees and costs**
  - *Klipsch Group, Inc. v. ePRO E-Commerce Ltd.*, 880 F.3d 620 (2d Cir. Jan. 25, 2018)[2]
- **Findings of failure to take reasonable steps to preserve in spoliation sanctions analyses**
  - *Paisley Park Enter., Inc. v. Boxill*, 330 F.R.D. 226, (D. Minn. Mar. 5, 2019)[3]
  - *Cruz v. G-Star Inc.*, 2019 WL 2521299 (S.D.N.Y. June 19, 2019), modified by *Cruz v. G-Star Inc.*, 2019 WL 4805765 (S.D.N.Y. Sept. 30, 2019)[4]

From not covering the right materials, to not covering the right people, to not notifying the right third-party service providers, to not rolling the hold out properly – or at all, it's clear that legal holds remain a minefield for litigants.

9.1.1 Hold On, I'm Coming

It's true that legal holds do not preserve data themselves, but they are the critical first step in the preservation process, ensuring that materials survive in situ long enough for you and your team to go get them. You are literally saying to everyone – just as Sam & Dave sang in 1966: "Hold On, I'm Coming"[5]

But, as the examples above make clear, this is easier sung than done effectively. Today's challenges include diversifying sources and source types (e.g., social media, collaboration tools), evolving custodian behavior (e.g., personal cloud storage, OTT messaging apps), and ever-increasing expectations (e.g., duties of competence extended to technology, new sources treated like old ones by judges).

[1] *United States v. Volkswagen AG*, No. 16-CR-20394 (USDC EDMI Jan. 11, 2017) (Third Superseding Information ¶¶ 22-25, available at https://www.justice.gov/usao-edmi/page/file/930021/download, and Plea Agreement at ¶¶ 73-82, available at https://www.justice.gov/usao-edmi/page/file/930026/download).
[2] *Klipsch Group, Inc. v. ePRO E-Commerce Ltd.*, 880 F.3d 620 (2d Cir. Jan. 25, 2018), available at https://casetext.com/case/klipsch-grp-inc-v-epro-e-commerce-ltd-1.
[3] *Paisley Park Enter., Inc. v. Boxill*, 330 F.R.D. 226, (D. Minn. Mar. 5, 2019), available at https://casetext.com/case/paisley-park-enters-inc-v-george-ian-boxill-rogue-music-alliance-llc-1.
[4] *Cruz v. G-Star Inc.*, 2019 WL 2521299 (S.D.N.Y. June 19, 2019), available at https://app.ediscoveryassistant.com/case_law/24556-cruz-v-g-star-inc, modified by *Cruz v. G-Star Inc.*, 2019 WL 4805765 (S.D.N.Y. Sept. 30, 2019), available at https:// casetext.com/case/cruz-v-g-star-inc-1.
[5] Sam & Dave, "Hold On, I'm Coming" (1966), available at https://www.youtube.com/watch?v=AREppyQf5uw.

9.2 WHAT MUST YOU PRESERVE, AND WHEN?

The Duty of Preservation

The duty of preservation is a foundational concept in our legal system that grows out of the common law concept of spoliation, which is more than 200 years old[6]:

- If courts exist to make determinations about disputed facts, and
- If the trier of fact must make those determinations using the available evidence,
  - Then, no litigant should be allowed to gain advantage in those determinations by destroying relevant evidence before the trier of fact can consider it

Additional discussion of the common law history of spoliation and preservation concepts is available in The Sedona Conference Commentary on Legal Holds, Second Edition: The Trigger & The Process.[7]

Although this common law duty of preservation is not directly codified in the Federal Rules of Civil Procedure, it is dictated by implication in Rule 26,[8] Rule 34,[9] and Rule 45.[10] Together, these three rules define the potential scope of discovery for litigants and third parties, and anything the rules may require you to produce is, inherently, something you need to preserve.

So, what is the scope defined by those rules?

9.2.1 The Scope of the Duty

The scope of potential discovery – and, therefore, of the duty to preserve – is deliberately broad, which is consistent with our court system's emphasis on truth-seeking over gamesmanship. As stated in one decision[11] involving discovery sanctions:

> Litigation is not a game. It is the time-honored method of seeking the truth, finding the truth, and doing justice. When a corporation and its counsel refuse to produce directly relevant information an opposing party is entitled to receive, they have abandoned these basic principles in favor of their own interests.

In its simplest form, the scope of discovery and preservation for ESI has four elements:

1. Documents
2. In your possession, custody, or control
3. That are potentially relevant
4. And unique

Documents

The definition of "documents" provided by the Federal Rules of Civil Procedure is expansive enough to encompass almost any sort of material in any format. Rule 34(a)(1)(A)[12] states that it covers "documents and electronically stored information – including":

[6] *Armory v Delamirie*, [1722] EWHC KB J94 (31 July 1722), available at https://www.bailii.org/ew/cases/EWHC/1722/J94.html.
[7] The Sedona Conference, *Commentary on Legal Holds, Second Edition: The Trigger & The Process*, 20 Sedona Conf. J. 341 (2019), available at https://thesedonaconference.org/publication/Commentary_on_Legal_Holds. 8 Fed R. Civ. P. 26, available at https://www.law.cornell.edu/rules/frcp/rule_26.
[9] Fed. R. Civ. P. 34, available at https://www.law.cornell.edu/rules/frcp/rule_34.
[10] Fed. R. Civ. P. 45, available at https://www.law.cornell.edu/rules/frcp/rule_45
[11] *Haeger v. Goodyear Tire & Rubber Co.*, 813 F.3d 1233 (9th Cir. 2016), available at https://casetext.com/case/haeger-v-goodyear-tire-rubber-co-3.
[12] Fed. R. Civ. P. 34(a)(1)(A), available at https://www.law.cornell.edu/rules/frcp/rule_34.

> . . . 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 . . .

The Committee Notes[13] on the rule emphasize the broadness again:

> The rule covers – either as documents or as electronically stored information – information "stored in any medium," to encompass future developments in computer technology. Rule 34(a)(1) is intended to be broad enough to cover all current types of computer-based information, and flexible enough to encompass future changes and developments.
>
> References elsewhere in the rules to "electronically stored information" should be understood to invoke this expansive approach.

Thus, nothing can be overlooked based purely on its format or source type; everything is potentially subject to the duty.

Possession, Custody, or Control

In addition to defining the broad scope of "documents," Rule 34(a)(1)[14] also specifies that the scope of discovery and preservation extends to those documents within "the responding party's possession, custody, or control." This phrase means that you are responsible, not just for the materials you physically or electronically possess, but for any that you legally control. Thus, materials maintained by third parties on your behalf are treated the same way as the records you actually possess yourself. If you have the right (or, in some cases, the ability) to obtain it, you are responsible for preserving and producing it.

Unfortunately for parties, there is some variation from jurisdiction to jurisdiction in exactly how far "possession, custody, or control" is deemed to extend. The three common standards – "Legal Right," "Legal Right Plus Notification," and "Practical Ability" – and their areas of applicability are broken down in detail in The Sedona Conference Commentary on Rule 34 and Rule 45 Possession, Custody, Or Control.[15]

Potentially Relevant

Among the "documents" that are in your "possession, custody, or control," the ones that may be discovered and must be preserved are those that are relevant.

Relevance is defined broadly by Federal Rule of Evidence 401.[16] That rule dictates that evidence is relevant if "it has any tendency to make a fact more or less probable than it would be without the evidence" and "the fact is of consequence in determining the action." The Committee Notes[17] to the rule state explicitly that this is an intentionally low bar because "[a]ny more stringent requirement is unworkable and unrealistic."

Thus, any documents in your possession, custody, or control that have any tendency to make any fact of consequence more or less likely are relevant, potentially discoverable, and required to be preserved.

[13]Fed. R. Civ. P. 34 advisory committee's note, available at https://www.law.cornell.edu/rules/frcp/rule_34.
[14]Fed. R. Civ. P. 34(a)(1), available at https://www.law.cornell.edu/rules/frcp/rule_34.
[15]*The Sedona Conference, The Sedona Conference Commentary on Rule 34 and Rule 45 "Possession, Custody, or Control,"* 17 Sedona Conf. J. 468, 482 (2016), available at https://thesedonaconference.org/publication/Commentary_on_ Rule_34_and_Rule_45_Possession_Custody_or_Control.
[16]Fed. R. Evid. 401, available at https://www.law.cornell.edu/rules/fre/rule_401.
[17]Fed. R. Evid. 401 advisory committee's note, available at https://www.law.cornell.edu/rules/fre/rule_401.

Unique

Finally, the scope of potential discovery and required preservation is limited to materials meeting the above criteria that are also unique. As specified by Rule 26(b) (2)(C),[18] discovery is not meant to be "unreasonably cumulative or duplicative." For ESI in particular, this is important, as it is in the nature of electronic systems to create numerous identical copies of materials, both for operation and for backup. Generally, there will be no additional evidentiary value to preserving numerous identical copies of the same materials.

9.2.2 Other Limitations

Beyond those four elements, there are two additional potential limitations on the scope of discovery that are less relevant to the question of preservation scope:

- First, as specified in Rule 26(b)(1),[19] the scope of discovery is limited to that which is "proportional to the needs of the case." Because any disputes over proportionality cannot be identified and resolved by the court until the matter is already underway, parties should not be quick to assume disproportionality and skip preservation.
- Second, as specified in Rule 26(b)(2)(B),[20] "[a] party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because of undue burden or cost." This is another type of proportionality requirement, specifically for electronically stored information, which recognizes that data recovery from some obsolete or challenging systems can be costly and burdensome. As with the general proportionality requirement, any disputes over proportionality cannot be identified and resolved by the court until the matter is already underway.

Preservation can always be stopped if it's later determined to be unnecessary, but lost data can never be recovered if it's later determined to have been necessary after all.

9.2.3 Triggers for the Duty

The duty to preserve documents often arises before a case is actually filed or commenced, because the duty arises not when there is litigation but when there is reasonable anticipation of litigation (or agency action, etc.). As explained in "Guideline 1" of The Sedona Conference Commentary on Legal Holds[21]:

> A reasonable anticipation of litigation arises when an organization is on notice of a credible probability that it will become involved in litigation, seriously contemplates initiating litigation, or when it takes specific actions to commence litigation.

[18]Fed R. Civ. P. 26(b)(2)(C), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[19]Fed R. Civ. P. 26(b)(1), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[20]Fed R. Civ. P. 26(b)(2)(B), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[21]The Sedona Conference, *Commentary on Legal Holds, Second Edition: The Trigger & The Process*, 20 Sedona Conf. J. 341 (2019), available at https://thesedonaconference.org/publication/Commentary_on_Legal_Holds.

Examples of triggering events include discovery of a legal or regulatory violation by an employee, receipt of a legal hold notice from a regulatory agency, hearing a terminated employee threaten suit, receipt of an actual complaint or subpoena, and many more.

9.3 WHAT SHOULD BE IN THE HOLD?

9.3.1 Evolution of Expectations

Formal, written legal holds became the focus of much attention in eDiscovery after the Zubulake V[22] ruling in 2004, in which a party was sanctioned for failing to issue a hold or take other necessary steps to ensure the preservation of relevant materials. In subsequent years,[23] this decision was cited in numerous others, and written legal holds became central to an effective eDiscovery preservation process.

For a time, the failure to issue a written legal hold was treated as per se gross negligence.[24] That absolute requirement for a hold in writing was softened by subsequent cases, however, which allowed for the possibility of circumstances in which oral holds or other approaches to preservation may be appropriate. See, e.g., *Chin v. Port Auth.* of N.Y. & N.J., 685 F. 3d 135 (2nd. Cir. July 10, 2012).[25]

9.3.2 Six Essential Elements of Holds

Despite the allowances for such circumstances, the issuance of a written legal hold (whether in paper or via email) is still considered best practice and the standard first step in any preservation process. Those legal holds can take a variety of forms and include a variety of optional content.

At root, though, all written legal holds should contain six essential elements. Each of these elements needs to be explained clearly and specifically:

1. First, the written hold should explain the legal obligations associated with the hold. This should include some explanation of the duty to preserve, the legal consequences for the organization if it is not fulfilled, and any internal consequences for employees who violate it. It is often helpful to point out that a request for individuals to preserve materials is a common legal step and not an indication that recipients are in any trouble.
2. Second, the written hold should explain the substantive scope of what must be preserved. This may include describing the underlying events, the relevant individuals inside and outside the organization with whom communication may have taken place, and more. This should also include the applicable time range, if any, and whether the hold applies going forward to newly created materials as well.
3. Third, the written hold should explain the types of materials that need to be preserved. This should include lists of relevant devices (e.g., laptops, phones, thumb drives), of relevant file types (e.g., email, spreadsheets, text messages), and of expected kinds of documents (e.g., internal financial reports, deal negotiation messages, annotated contract drafts, etc.).

[22] *Zubulake v. UBS Warburg LLC*, 229 F.R.D. 422 (S.D.N.Y. 2004), available at https://casetext.com/case/zubulake-v-ubs-warburg-llc-3.
[23] Victor Li, "Looking back on Zubulake, 10 Years Later," ABA Journal (Sept. 1, 2014), available at http://www.abajournal.com/magazine/article/looking_back_on_zubulake_10_years_later.
[24] Rachel S. Fendell, *Impact Of Chin Decision On Pension Committee, Mondaq*, https://www.mondaq.com/unitedstates/disclosure-electronic-discovery-privilege/190306/impact-of-chin-decision-on-pension-committee (Aug. 6, 2012).
[25] *Chin v. Port Auth. of N.Y. & N.J.*, 685 F. 3d 135 (2nd. Cir. July 10, 2012), available at https://scholar.google.com/scholar_case?case=11269039069845908318.

4. Fourth, the written hold should explain the process that will be used for preserving and collecting the subject materials. These are the specific instructions the recipients of the hold are to follow for handling the materials they possess that are subject to the hold. Should they preserve them in place? Segregate them in some way? Take other steps? When and how will they be contacted about collection of those materials?
5. Fifth, the written hold should explain how and with whom the recipients may communicate about the hold. This should include both any prohibitions on communication about the hold or the underlying matter with peers, as well as instructions for who should be contacted with any questions about scope or process. This is especially important if you wish to treat the hold as a privileged communication.
6. Sixth and finally, the written hold should request some type of confirmation from the recipient that they have received the hold, reviewed the hold, and will abide by the hold. This confirmation of receipt and compliance may take the form of a sheet that is signed and returned, an email response, on online form, or some other mechanism.

It is important to remember that the hold must cover not only the devices and materials of individual custodians, but also departmental and enterprise systems and any automated janitorial functions that may be running on them. We will discuss this further below.

9.3.3 Other Components to Consider

In addition to the essential elements described above, there are a variety of optional elements you can include to accomplish more with the distribution of your legal hold. The two most commonly included additions are:

Custodian Surveys for Collection

Many organizations also use the distribution of the written legal hold as an opportunity to begin gathering details for collection planning. They distribute some form of custodian questionnaire with the hold and require its completion as well. These may be created as paper questionnaires, electronic forms, or online surveys, and they can take the place of initial interviews for many custodians.

Frequently Asked Questions

Employees of an organization who have not been through a legal hold process before typically have little familiarity the process or its role in discovery and litigation. Questions about it are common, as are questions about scope and process. To aid employees in their understanding, many organizations draft an FAQ (Frequently Asked Questions) for distribution with the hold.

This FAQ typically restates much of the information from the hold in a less formal way and attempts to anticipate and answer the likely questions about context, scope, and process.

Six Elements

Written legal holds should generally contain six essential elements: legal obligations, substantive scope, materials to be preserved, preservation process, communication instructions, and a compliance confirmation.

9.4 LEGAL HOLD PROCESSES AND POLICIES

9.4.1 Documentation, Consistency, and Defensibility

To consistently execute effective legal holds, there are five key activities for which reliable processes need to be in place, and if possible, they should each also be addressed by a written hold policy. Consistency is key to defensibility, and documentation is key to consistency. As The Sedona Conference Commentary on Legal Holds[26] states clearly in its guidelines:

> **Guideline 2**
>
> Adopting and consistently following a policy or practice governing an organization's preservation obligations are factors that may demonstrate reasonableness and good faith.
>
> **Guideline 9**
>
> An organization should consider documenting the procedure of implementing the legal hold in a specific case when appropriate.

The five key activities to address in this way are hold initiation, hold drafting, recipient identification, compliance monitoring, and hold release.

9.4.2 Initiation: The Holding Hour Is Upon Us

We have already discussed the potential range of triggering events for the duty to preserve, but what happens when one occurs? How does word filter to the appropriate individual? Who is the individual responsible for taking action? What actions do they take to initiate the process? Are the initial steps internal, or executed with outside counsel? On what timeline do they act?

Establishing a consistent, reliable process and policy for hold initiation requires that each of these questions be addressed. Many organizations establish three components to address this activity:

1. An organizational policy dictating when legal holds should be implemented
2. A legal department process for initiating legal hold creation and issuance
3. An employee handbook policy describing employees' duty to report certain incidents

Together, these three components can go a long way towards demonstrating reasonableness and good faith in your efforts. As The Sedona Conference Commentary on Legal Holds[27] states in Guideline 3: "Adopting a procedure for reporting information relating to possible litigation to a responsible decision maker may assist in demonstrating reasonableness and good faith."

9.4.3 Drafting: Get It in Writing

The next activity that can benefit greatly from a standardized process and a documented policy is the drafting of the hold to be issued. Once the responsible individual has identified a triggering event and started down the road to hold issuance, who will actually be responsible for drafting? Who will contribute to the legal substance? Who will address technical questions about subject source types or affected enterprise systems? Who will address any cross-border or data privacy concerns?

[26]The Sedona Conference, *Commentary on Legal Holds, Second Edition: The Trigger & The Process*, 20 Sedona Conf. J. 341 (2019), available at https://thesedonaconference.org/publication/Commentary_on_Legal_Holds.
[27]The Sedona Conference, *Commentary on Legal Holds, Second Edition: The Trigger & The Process*, 20 Sedona Conf. J. 341 (2019), available at https://thesedonaconference.org/publication/Commentary_on_Legal_Holds.

Ultimate responsibility for the contents of the legal hold typically rests with an organization's general counsel and the lead outside counsel for the matter, which is typically documented in the organization's overall legal hold policy. Additionally, it is common for a legal department to have a documented process for drafting to help ensure consistency and completeness. Common topics addressed include:

- People to Involve
  - Inside and outside counsel, IT/ IS, RM/ KM, compliance, data protection officer, discovery service providers, etc.
- Potential Sources to Flag
  - Potential custodian devices and files, departmental systems and files, enterprise systems, backup systems, third-party service providers, etc.
- Collection Methods to Specify
  - Techniques and providers approved for organizational use from which to choose
- Other Potential Issues to Consider
  - Cross-border implications, data privacy implications, employee turnover issues
- Standardized Templates to Use
- Legally vetted templates for a standard hold and for any recurrent matter types

Having a consistent process that includes input from the right individuals, consideration of all common issues, and the use of predefined approaches and templates can go a long way towards both ensuring effectiveness and demonstrating reasonableness and good faith in your efforts.

9.4.4 Identification: All Key Players, Please Step Forward

The next activity for which a repeatable process is essential is identification of the appropriate hold recipients. Even a timely and well-written hold will not be effective if it does not reach everyone it needs to reach. So, who does it need to reach? Does it need to go to the entire company? To a particular department? To specific individuals? What about executive management? Who is responsible for relevant enterprise information systems? Are there outside, third-party recipients that need to be added too?

Since Zubulake V[28] in 2004, the phrase "key players" has been used to describe the essential recipients of a legal hold within an organization. Key players are those with the direct knowledge of the underlying events or those most likely to have relevant information or materials. This is often, but not always, managers and executives. Even plant-level employees have been deemed key players when they had relevant knowledge.[29]

[28] *Zubulake v. UBS Warburg LLC, 229 F.R.D.* 422 (S.D.N.Y. 2004), available at https://casetext.com/case/zubulake-v-ubs-warburg-llc-3.
[29] *See, e.g., Consolidated Aluminum Corp. v. Alcoa, Inc.*, 244 F.R.D. 335 (M.D. La. 2006), available at https://casetext.com/case/consolidated-aluminum-corporation-v-alcoa

It is not always possible to identify all key players in the abstract; you may need to communicate with some or all of the key players and ask for referrals to others. Depending on the sequence of events, this may mean sending the hold to additional recipients after the initial distribution, as you learn new details. Other common pitfalls include:

- **Enterprise and departmental systems** – organizations may have any number of enterprise systems (e.g., email, backup, or document management) and departmental systems (e.g., benefits, payroll, research, or compliance), each with different owners and their own automated janitorial functions (or tape recycling schedules) continually deleting older files, which will need to be suspended if any relevant materials are at risk
    - Each owner responsible for systems containing relevant information will need to be a recipient of the hold to ensure those mechanisms of deletion are halted
- **Third-party providers** – organizations very commonly outsource one or more business functions, like payroll or benefits (or even email), to specialized third-party providers, and the data they possess on your organization's behalf is subject to the same duty to preserve, as we discussed above
    - Third-party service providers in possession of potentially-relevant materials will need to be recipients of the legal hold as well, and your service contracts with them may specify particular notice procedures to follow for each provider
- **Employee turnover** – employee departure is a common occurrence in organizations of almost any size, and many organizations wipe and reissue employee devices when that happens (and deactivate email accounts, CRM accounts, etc.), which is a problem if the individual was subject to a legal hold and unique, relevant materials are lost in the process
    - Whoever in HR or IT typically handles these steps also needs to be a recipient of the hold and be generally kept informed about active holds that should change normal device recycling steps, account deactivation steps, etc.

9.4.5 Monitoring: Once Is Never Enough

Ongoing compliance monitoring after hold issuance is the activity for which it is most important to have a consistent, documented process. As has been made clear in case after case, failure to check if individuals are actually complying, or failure to remind them as needed, can be just as consequential as failure to issue the hold in the first place.[30]

Common steps to ensure ongoing compliance with the hold include:

- **Receipt and compliance verification** – having employees sign a document or electronic form, or send an email, confirming that they have received the hold, understood the hold, and will comply with the hold; this is typically covered as part of the hold itself, as discussed above
    - The same can be applied to those responsible for suspending janitorial functions on enterprise or departmental systems, including backup systems and tapes

[30] *See, e.g., Pension Committee of University of Montreal Pension Plan v. Banc of America Securities*, 685 F. Supp. 2d 456 (S.D.N.Y. Jan 15, 2010), available at https://www.courtlistener.com/opinion/1881971/univ-of-montreal-pension-plan-v- banc-29 of-am-sec/; *Chin v. Port Auth. of N.Y. & N.J.*, 685 F. 3d 135 (2nd. Cir. July 10, 2012), available at https://scholar.google.com/scholar_case?case=11269039069845908318.

- **Spot checking** – it also advisable to establish a regular schedule for checking in with at least a sampling of the subject custodians (checking everyone may not be feasible) to check that they are in fact complying and materials are being preserved
- **Reissuance** – since legal matters and the holds associated with them can continue for months or years, it is also advisable to establish a schedule for periodic reissuance of the hold as a reminder to those it covers (quarterly is common); the specific scope of the hold may also need to be revised as a legal matter evolves and more is learned

As we noted at the beginning, a legal hold is not itself preservation, and if it is not followed by the other steps necessary to ensure actual preservation takes place – like ongoing compliance monitoring, then whether or not a hold was issued doesn't really matter.

9.4.6 Release: When All Is Said and Done

Finally, it is valuable to have a standardized process, documented in a policy, for the release of legal holds after a matter has concluded. The Sedona Conference Commentary on Legal Holds[31] includes this in Guideline 11:

> Any legal hold process should include provisions for releasing the hold upon the termination of the duty to preserve, so that the organization can resume adherence to policies for managing information through its useful life cycle in the absence of a legal hold.

Having a consistent standard and a defined process for the review of when to release a legal hold can do a lot to demonstrate good faith in your preservation efforts. The three key factors to consider before deciding to release a legal hold are:

1. Whether there are any remaining court or agency orders requiring retention
2. Whether there is a possibility of related future litigation (e.g., appeal, new suit)
3. Whether the materials being preserved are potentially relevant to any other matters

If there are no applicable orders, no reasonably foreseeable future litigation, and no reason to preserve for other matters, then a hold can be released and an organization can revert to its default management, retention, destruction, and recycling policies for documents and devices.

Trust But Verify

> Ongoing compliance monitoring is the activity for which it is most important to have a documented process. Failure to monitor compliance, can be just as consequential as failure to issue a hold in the first place.

[31] *The Sedona Conference, Commentary on Legal Holds, Second Edition: The Trigger & The Process, 20 Sedona Conf. J. 341 (2019), available at https://thesedonaconference.org/publication/Commentary_on_Legal_Holds.*

9.5 METHODS FOR DISTRIBUTING LEGAL HOLDS

9.5.1 Paper

The first and simplest tool available for legal holds is, of course, paper. Before the advent of the options discussed below, paper holds, signed paper confirmations, and paper reminders were the norm, and for smaller organizations (e.g., those in a single office location), paper may still be a good choice. It is simple and inexpensive to create, distribute, and document holds in this fashion. For larger or more geographically- distributed organizations, paper can quickly become logistically cumbersome and time consuming, however.

9.5.2 Email

As consistent, universal email usage became typical, the same processes were executed using messages in the body of emails – holds distributed as emails, confirmation responses done as reply emails, etc. This is also a suitable approach for small or medium organizations, and email lets you easily extend the simple approach of paper beyond a single office location. For larger organizations, however, manually tracking the number of emails back and forth that will be required can become just as logistically cumbersome as distributing and collecting paper.

9.5.3 Electronic Forms

For medium or large organizations, it is now common to create and use electronic forms rather than just paper or emails. These are forms with defined fields that allow recipients to electronically "sign" the forms and then fill out any other requested information (e.g., preliminary custodian survey information). Standard field entries make the aggregation and tracking of the responses much easier than it is with loose paper or emails. These types of forms are most often created as Adobe PDF files[32] or as Microsoft Excel files.[33] Each has advantages and disadvantages, but the current trend seems to be toward PDF forms, which are arguably easier to build and which look more like traditional paper forms to recipients.

9.5.4 Purpose-Built Tools

Today, there are also a variety of purpose-built tools specifically for creating, distributing, and managing legal holds. There are more than a dozen offerings of this type in the marketplace, and more are sure to appear as the industry continues to grow. Some of these are standalone applications, some are SaaS solutions, and some are modules integrated into larger litigation management or eDiscovery software suites. All of them provide a measure of automation and standardization for the creation, distribution, tracking, and refreshing activities, allowing an organization to centrally manage the numerous simultaneous holds common to large organizations.

Scale to Fit

For mid-size organizations, it is common to use electronic forms rather than paper or emails, and for large organizations, a variety of purpose-built tools are available.

[32]"Create and distribute PDF forms," Adobe Support (Sept. 7, 2022), available at https://helpx.adobe.com/acrobat/using/creating-distributing-pdf-forms.html.
[33]"Overview of forms, Form controls, and ActiveX controls on a worksheet," Microsoft Support, available at https://support.office.com/en-us/article/Overview-of-forms-Form-controls-and-ActiveX-controls-on-a-worksheet-15ba7e28-8d7f-42ab- 9470-ffb9ab94e7c2.

9.6 WHAT ABOUT PRIVILEGE?

In most situations, legal hold notices are communications from an in-house counsel or an outside counsel to employees of an organization, which brief them on a legal situation and the need to hold materials for it. As such, legal hold notices are typically considered both privileged attorney-client communications (because they are the communication of legal guidance) and protected attorney work product (because they reveal the attorney's thinking about the matter).

This general principle can be seen applied in numerous cases. For example, *Gibson v. Ford Motor Co.*, 510 F. Supp. 2d 1116 (N.D. Ga. 2007)[34] includes the following passage discussing a request for the production of legal hold notices issued by the Defendants:

> Plaintiffs request the document sent to Defendant's employees instructing them not to destruct certain kinds of documents required to be maintained as a result of this litigation. . . . In the Court's experience, these instructions are often, if not always, drafted by counsel, involve their work product, are often overly inclusive, and the documents they list do not necessarily bear a reasonable relationship to the issues in litigation. This is not a document relating to the Defendant's business. Rather, the document relates exclusively to this litigation, was apparently created after this dispute arose, and exists for the sole purpose of assuring compliance with discovery that may be required in this litigation. Not only is the document likely to constitute attorney work-product, but its compelled production could dissuade other businesses from issuing such instructions in the event of litigation. Instructions like the one that appears to have been issued here insure the availability of information during litigation. Parties should be encouraged, not discouraged, to issue such directives. Defendants are not required to produce these materials.

“Not only is the document likely to constitute attorney work-product, but its compelled production could dissuade other businesses from issuing such instructions in the event of litigation.

In some situations, these protections may not be afforded, however. For example: if the hold notice is sent by a non-lawyer executive rather than counsel; if the hold notice specifies that it is not confidential and should be shared with co-workers; or, if spoliation has taken place, requiring further discovery about the reasonableness of preservation efforts.

9.7 EVOLUTION IS ENDLESS

We have touched a few times on the diverse range of potential sources that now exist and that must be considered for coverage by the hold. In addition to remembering to think about newer source types like collaboration tools and generative AI applications, the deliberately-expansive definition of “documents” used by the rules means that you must also account for the fact that technology and your custodians' use of it is constantly evolving.

[34] *Gibson v. Ford Motor Co.*, 510 F. Supp. 2d 1116 (N.D. Ga. 2007), available at https://scholar.google.com/scholar_case?case=17383954170793254828.

Because of this reality, your list of sources and source types to consider must evolve over time too. Your documented processes should include periodic review of your potential source lists (e.g., annually) to see if they need to be updated with newly acquired enterprise tools, new kinds of employee devices, or emerging technologies (e.g., apps, cloud services) being adopted by your custodians. To do this effectively, you will need to consult with your enterprise IT resources, who can provide updates on the organization, and your forensic collection service providers, who can provide updates on global usage trends and evolving industry expectations.

9.7 KEY TAKEAWAYS

There are eight key takeaways from this chapter to remember:

1. **Failures to implement effective legal holds can carry serious consequences**. Failure to implement an effective legal hold can lead to determinations that reasonable steps to preserve were not taken, to adverse inferences and other sanctions, or even to criminal obstruction charges.
2. **The scope extends to all potentially-relevant documents within your control.** The scope of the duty of preservation extends to all unique, potentially-relevant documents or ESI – of any type – in your possession, custody, or control (which includes materials held by third-party service providers).
3. **The duty is triggered whenever there is a reasonable anticipation of litigation.** The duty of preservation is triggered whenever there is a reasonable anticipation of litigation (or agency action, etc.), which can happen well before a case is filed.
4. **There are six essential elements that should be included in a legal hold.** An effective legal hold should include information regarding: (1) the legal obligations associated with the hold; (2) the substantive scope of what must be preserved; (3) the types of materials that must be preserved; (4) the process that is to be used for preservation and collection; (5) how and with whom recipients may communicate about the hold; and, (6) a receipt and compliance confirmation mechanism. Additionally, you may consider including custodian surveys for collection or
5. **Documented, consistent processes are more reliable and more defensible.** Consistent processes are more defensible than ad hoc ones, and documented processes are more defensible still. Ideally, both overall written policies and project-by-project process documentation should be created and maintained.
6. **There are five core hold activities for which such processes should be developed.** Hold initiation, hold drafting, recipient identification, compliance monitoring, and hold release are the key activities for which consistent, documented processes should be created. In particular, ongoing monitoring of hold compliance is crucial to success.
7. **Available tools include paper, email, electronic forms, and purpose-built tools.** Your options for hold implementation tools range from paper to purpose-built software, and the right choice will depend heavily on the size of your organization, the size (or number) of your matter(s), and other situation-specific considerations (e.g., compatibility with existing enterprise systems or eDiscovery tools).
8. **Technology is always evolving, and we must evolve with it.** Your lists and plans must be periodically reviewed and updated to reflect new enterprise tools, new employee devices, and new communication technologies in the marketplace.

Chapter 10

Beyond the Four Corners: Evolving Electronic Documents

About this Chapter

In this chapter, we will review key issues for practitioners to consider regarding the challenges created by new and evolving ESI sources. We will begin with current challenges created by the evolving nature of electronic "documents" that cut across source types, and then we will discuss some source-specific issues of which practitioners should be aware.

10.1 THE ONLY CONSTANT IS CHANGE

Identification and preservation are the first and most fundamental phases of an electronic discovery effort. Almost every other type of discovery process failure can be fixed with adequate time and money, but once unique, relevant electronically-stored information (ESI) is gone, it's gone. Unfortunately, the challenges of identifying, preserving, and collecting relevant ESI continue to grow as old sources evolve, new sources emerge, and the behaviors of organizations and individuals adapt.

10.1.1 Evolution of Sources and Behavior

The story of the past decade has been one of the long, slow march into the cloud, as organizations have transitioned to new software-as-a-service solutions and individuals have transitioned to new messaging and collaboration tools. One of the clearest illustrations of this trend is the adoption and evolution of the Microsoft 365 offering and the Microsoft Teams application.

Microsoft 365

In 2011, Microsoft launched a new, cloud-based subscription service called Office 365 consisting of their Microsoft Office productivity applications. Office 365 allowed employees and organizations to utilize up- to-date online versions of Microsoft Office applications like Outlook, Excel, Word, and PowerPoint to create their business communications and documents.

Adoption was rapid, and personal, educational, and small business licensing packages followed. By October 2019, Microsoft surpassed 200 million commercial monthly active users.[1] As adoption increased, the range of included applications and capabilities expanded, and by early 2020, Microsoft transitioned to the broader name Microsoft 365 (M365).

A major turning point came in 2020, when the pandemic and the accompanying transition to remote or hybrid work accelerated adoption and usage even more. As of April 2022, Microsoft reported 345 million paid seats for Microsoft 365 and a 17% year- over-year increase in revenue from Microsoft 365.[2] It was also reported that it was in use by over a million companies worldwide,[3] over 870,000 of which were in the United States. As of September 2022, Microsoft 365 customers were adding "over 100 petabytes of new content each month."[4]

Microsoft Teams

As part of the expansion of what would eventually be called M365, Microsoft added in 2017 a new chat and collaboration application called Teams. Teams was created to compete with Slack, which was a self-described "digital HQ" merging communications, information, and documents into a single collaboration application. After its launch in 2013, Slack was the fastest-growing workplace software ever, topping 500,000 daily users in 2015.[5] Despite Slack's four- year head start, however, Microsoft Teams quickly surpassed Slack, reaching 20 million daily users by November 2019.[6]

[1]Mary Jo Foley, "A new Microsoft cloud category to watch: The Microsoft 365 number," ZDNET (Oct. 23, 2019), available at https://www.zdnet.com/home-and-office/work-life/that-big-microsoft-365-teams-and-outlook-outage-heres-what- went-wrong/.
[2]Tony Redmond, "Office 365 Reaches 345 Million Paid Seats," Office 365 for IT Pros (Apr. 28, 2022), available at https://office365itpros.com/2022/04/28/office-365-number-of-users/.
[3]Lionel Sujay Vailshery, "Number of Office 365 company users worldwide 2022, by country," Statista, (Feb. 23, 2022), available at https://www.statista.com/statistics/983321/worldwide-office-365-user-numbers-by-country/.
[4]Omar Shahine, "Microsoft is recognized as a Leader in the 2021 Forrester Wave for Content Platforms," Microsoft 365 Blog (Sept. 28, 2021), available at https://www.microsoft.com/en-us/microsoft-365/blog/2021/09/28/microsoft-is-rec- ognized-as-a-leader-in-the-2021-forrester-wave-for-content-platforms/.
[5]Jack Linshi, "This 1-Year-Old Startup Says It's the Fastest-Growing Business App Ever," Time (Feb. 12, 2015), available at https://time.com/3705218/slack-business-app/.
[6]Mary Jo Foley, "Microsoft says it has 20 million daily active Teams users," ZDNET (Nov. 19, 2019), available at https://www.zdnet.com/article/microsoft-says-it-has-20-million-daily-active-teams-users/.

This rapid growth was then turbocharged by the pandemic and consequent shift to remote and hybrid work, resulting in geometric growth for Microsoft Teams. Daily active users tripled in 2020, and then they more than doubled again in 2021. By the end of 2022, Teams had over 270 million monthly active users.[7]

10.2 EVOLVING ELECTRONIC DOCUMENTS

The rapid changes described above have created four major types of preservation and collection challenges that cut across a variety of newer source types: linked and dynamic ESI, threaded message ESI, emojis in ESI, and ephemeral and encrypted ESI.

10.2.1 Linked and Dynamic ESI

As a result of the ongoing transition to more dynamic, cloud-based office and collaboration tools, the old conception of "documents" and "document families" as static things is giving way to a new paradigm in which documents are dynamic and family relationships are virtual.

Traditionally, a document was attached to an email or other message as a self-contained unit containing the data of both the parent email and the child attachment. The relationship between these two electronic documents was easy to preserve and collect. Now, so-called "modern attachments" have changed how this works.[8]

Modern attachments (or pointers, or links) are a feature of Microsoft OneDrive for Business that integrates with Microsoft Outlook clients and Teams clients, and versions of the same functionality are being implemented in other platforms such as Google Workspace and Adobe Cloud. Instead of attaching a copy of a document to an email message, the system inserts a link to the original document that a recipient can click to open it directly.

This new paradigm presents several open questions for electronic discovery:

- Should a linked document be treated the same way as a traditional attachment?
- What if the current version of the document is different than when the link was sent?
- Are parties required to try to recover or recreate past versions of the document?
- What is the right balance between burden and reasonableness in this context?

Courts have not yet reached a consensus on the question, with some ordering the production of linked documents and some declining to do so.[9] What the analyses so far have in common is a fact-specific focus that looks a lot like a traditional proportionality analysis. Factors considered

[7]Lionel Sujay Vailshery, "Microsoft Teams: number of daily active users 2019-2022," Statista (Jan 13, 2023), available at https://www.statista.com/statistics/1033742/worldwide-microsoft-teams-daily-and-monthly-users/.
[8]Staci Kaliner, Monica McCarroll, and Ben Barnes, "Let's Start by Calling Them What They Are for Discovery: 'Pointers' Not 'Modern Attachments,'" Legaltech News (Aug. 11, 2022), available at https://www.law.com/legaltechnews/2022/08/11/ lets-start-by-calling-them-what-they-are-for-discovery-pointers-not-modern-attachments/.
[9]See e.g., *Nichols v. Noom, Inc.*, 2021 WL 948646 (S.D.N.Y. Mar. 11, 2021), available at https://app.ediscoveryassistant.com/case_law/32615-nichols-v-noom-inc; IQVIA Inc. v. Veeva Systems, Inc., 2019 WL 3069203 (D.N.J. Jul. 11, 2019), available at https://app.ediscoveryassistant.com/case_law/24806-iqvia-inc-v-veeva-sys-inc.

have included the language of the discovery agreement between the parties, the scope of the specific request, the technical or financial burden of fulfilling the request, and the importance of the linked documents.

Beyond linked attachments, other types of dynamic content are also creating new challenges. For example, what is the right way to preserve, collect, and produce particular views of a dynamic dashboard?[10]

10.2.2 Threaded Message ESI

Traditionally, organizations treated enterprise email as foundational to their information management and eDiscovery programs and treated chat and messaging applications as casual and secondary. Today, a major transition is under way from email communication to chat and collaboration tool communication, and the younger an employee is, the more likely they are to prioritize these new communication channels over traditional email.

The proliferation of mobile device sources, social media sources, and collaboration tool sources has made message thread unitization a common question for eDiscovery. These source types frequently include ongoing threads of back-and-forth messages (e.g., WhatsApp text message threads, social media direct message threads, Slack channel threads, etc.), which can span extended periods of time. Although the specifics vary by source, these message threads are often maintained in ongoing logs that are not conducive to efficient review or later use as evidence. Rather than present weeks or months of messages in a single document, it is typical to unitize these logs into separate, shorter documents for review and production.

When doing so, some judgment must be exercised about what size the units should be. Individual messages stripped of thread context are also not ideal (as courts have pointed out[11]), so some middle ground between massive logs and single messages is preferred. It is common to unitize such materials into 24-hour chunks, so that each day's communications become a single document, but other divisions may be rational depending on your materials and case.

This unitization is typically performed during processing, prior to ECA, review, and production, but production implications should be considered when the best way to unitize and produce such materials.

10.2.3 Emojis in ESI

Over the past decade, adoption of emoji use has become widespread. By 2019, according to a survey report from Adobe,[12] more than 90% were using emojis in personal communications, and more than 60% were using emojis in work communications. For example, according to a Microsoft spokeswoman,[13] emoji use in 2019 was "basically universal" among the 13 million daily active users of Microsoft Teams.

As these communication channels have become more frequent discovery sources, so too have emojis shown up more frequently in cases. In 2019, Santa Clara University Professor of Law Eric Goldman published "Emojis and the Law"[14] in the Washington Law Review, which revealed that

[10] *Famulare v. Gannett Co., 2022* WL 815818 (D.N.J. Mar. 17, 2022), available at https://casetext.com/case/famulare-v-gannett-co.
[11] See, e.g., *Laub v. Horbaczewski*, 331 F.R.D. 516 (C.D. Cal. Apr. 22, 2019) (Magistrate Judge expressing a preference for "aggregated" formats preserving "the integrity of the threads of communication reflected in the text messages"), available at https://casetext.com/case/laub-v-horbaczewski.
[12] Adobe, *Emoji Trend Report 2019*, (Jul. 15, 2019), available at https://www.slideshare.net/adobe/adobe-emoji-trend-report-2019.
[13] Christopher Mims, *Yes, You Actually Should Be Using Emojis at Work*, WALL STREET JOURNAL (July 20, 2019), available at https://www.wsj.com/articles/yes-you-actually-should-be-using-emojis-at-work-11563595262.
[14] Eric Goldman, *Emojis and the Law*, 93 WASH. L. REV. 1227 (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3133412.

"[b]etween 2004 and 2019, there was an exponential rise in emoji and emoticon references in US court opinions."[15] The presence of these emojis creates special challenges for eDiscovery and litigation – both technical challenges and challenges of interpretation.

First, the volume and diversity of emojis make it a challenge for discovery tool developers and service providers to keep up with supporting them all. There are an enormous and growing number of emojism,[16] and they work in a variety of ways. The cross-platform emojis recognized by the Unicode Consortium exist as alphanumeric codes that various software knows to replace by displaying a corresponding image, while platform-specific and user-created emojis may be based on custom, platform-specific codes, or may exist only as image files that function more like attachments. This support problem even extends to the word processing software used to write briefs and opinions and to the search tools powering case law databases.

Second, there is a challenge associated with the contextual relationship between emojis and text when they are used together. If included emojis are not all captured and displayed, it can lead to material alterations to messages and their meaning. For example, a message might include an emoji indicating it was intended humorously or sarcastically. If that emoji is omitted during collection or not displayed during review, the message might appear misleadingly serious or literal. Communications using multiple emojis can also be very ambiguous. It may not be clear to you what a custodian was attempting to communicate or what a recipient understood.

10.2.4 Ephemeral and Encrypted ESI

Another change over the past decade has been the increase in availability and use of ephemeral messaging and end-to-end encryption. Ephemeral messaging allows for the automatic deletion of sent messages after a set amount of time. As far back as 2016, ephemeral messaging applications were being used by 56% of smartphone owners ages 18-29,[17] and in 2017, Uber made headlines[18] for its use of ephemeral messaging app Wickr, and they were not alone.[19]

Ephemeral messaging can have advantages for organizations, including reducing unnecessary data retention and increasing the security of sensitive communications. When it comes time for discovery or investigation, however, ephemeral messaging can create challenges. First, all automatic deletion of new relevant communications must be suspended, which may be difficult if central control of the relevant channels' settings is not possible. Second, a company's intentions may come under scrutiny – particularly in the absence of clear usage and retention policies – if there is some question as to why certain communications were being deleted.

Properly implemented end-to-end encryption, on the other hand, protects messages by ensuring it's impossible for anyone but the sender and recipient to read them. Availability and use of this has also been on the rise. For example, encrypted messaging app Signal has been one of the fastest-growing new communication platforms in the world. As of the end of 2021, it had more than 40 million monthly active users,[20] and many of them were using it for both private and professional communications.

As adoption and use has increased, so has the frequency with which Signal has been identified

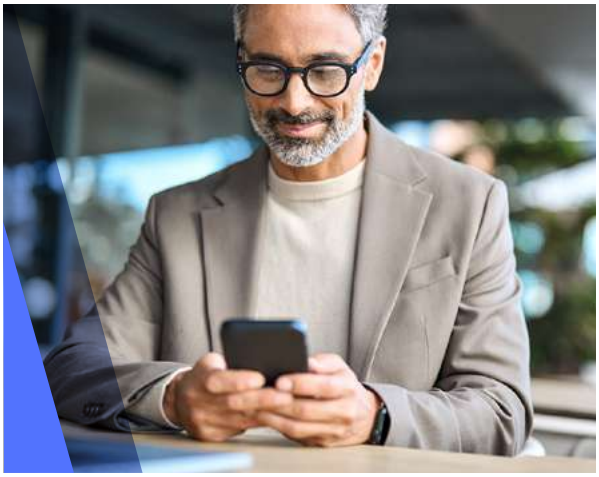
[15]Dami Lee, *Emoji are showing up in court cases exponentially, and courts aren't prepared*, THE VERGE, https://www.theverge.com/2019/2/18/18225231/emoji-emoticon-court-case-reference (Feb. 18, 2019).
[16]Today, there are more than 3,500 emojis recognized by the Unicode Consortium. Beyond those cross-platform emojis, many platforms also include platform-specific emojis or allow for the creation of custom emojis. In popular collaboration tool Slack, for example, "26 million custom emojis have been created since the feature was introduced."
[17]Greenwood, Perrin, & Duggan, *supra* note 16.
[18]Julie Bort, Uber's *CEO acknowledged his workers' use of secretive messaging apps — and says he banned them*, INSIDER, http://www.businessinsider.com/ubers-new-ceo-has-banned-secretive-messaging-apps-2017-11 (Nov. 29, 2017).
[19]Heather Kelly, *Secret message apps on the rise at work*, CNN BUSINESS, http://money.cnn.com/2017/12/11/technology/secret-messaging-apps-work/index.html (Dec. 11, 2017).
[20]David Curry, "Signal Revenue & Usage Statistics (2023)," Business of Apps (Jan. 9, 2023), available at https://www.businessofapps.com/data/signal-statistics/.

as a source of relevant communications that must be identified, preserved, and collected to fulfill litigation duties, compliance duties, and other recordkeeping obligations. Government agencies have recently placed special emphasis on the importance of organizations accounting for all communication channels in use by their employees, particularly encrypted or ephemeral channels like Signal.[21]

End-to-end encrypted applications like Signal pose a variety of challenges as discovery sources. Data is typically available only from the users' devices and only with the users' credentials. In some cases, it is only possible to collect the data by gaining root access to the mobile device or by using expensive specialized tools. Often, the best option is "collecting" via screen captures or screen recordings, which can then be run through optical character recognition and manually annotated with relevant metadata.

10.3 SOURCE-SPECIFIC ISSUES

Beyond the four cross-source issues discussed above, it is also important for practitioners to be aware of some source-specific challenges associated with the five sources most commonly associated with the issues above. They are mobile devices, OTT messaging apps, collaboration tools, Microsoft 365, and social media.



Mobile Devices

Mobile devices – smartphones in particular – have become ubiquitous for both personal and business life. Like all consumer technology, there are a plethora of models and types available, and new ones are released by each maker each year. And, because many organizations have adopted bring-your-own-device policies (BYOD), organizations may have a much wider variety of smartphones as potential sources than computers (which still tend to be organization-selected and issued).

Smartphones are more difficult, more costly, and more time-consuming to collect and process than computers. The difficulty, cost, and time can vary from model to model, from maker to maker, and from operating system to operating system. Collection directly from smartphones requires specialized tools like those used to collect from a custodian's computer. Collections instead from cloud-based backups of the smartphone in question are sometimes also an option.

Different models run different types of operating systems, and the operating systems differ in functionality and are updated regularly. Updates can affect the way in which applications store their data or how they are backed up. In other words, data that can be forensically extracted today, may not be able to be extracted tomorrow, or vice versa.

At a high level, applications that come pre-installed on a mobile device when you take it out of the box and power it on, such as Contacts, SMS, MMS, Calendar, Photos, and Video, will typically

[21]See, e.g., Matthew Goldstein and Emily Flitter, "Texting on Private Apps Costs Wall Street Firms $1.8 Billion in Fines," The New York Times (Sept. 27, 2022), available at https://www.nytimes.com/2022/09/27/business/banks-fined-texting-sec. html; U.S. Department of Justice, "Evaluation of Corporate Compliance Programs (Updated March 2023)" 17-18 (Mar. 3, 2023), available at https://www.justice.gov/criminal-fraud/page/file/937501/download.

be extracted from the handset during a standard imaging process using forensic tools. These applications are known as "stock" applications.

Additionally, it is important not to overlook less common mobile devices that may, at times, be relevant, such as vehicle GPS or data systems,[22] wearable devices like fitness trackers,[23] etc.

OTT Messaging Apps

Third-party applications on mobile devices, which are applications that the user downloads onto the handset from digital storefronts like the Apple App Store or the Google Play Store, may or may not be extracted from the handset during a standard device collection process. As noted above, this may be because of end-to-end encryption or other security measures implemented by the app's developers.

This varies not just from app-to-app but even across devices and operating systems. For example, WhatsApp data is stored in an encrypted format on recent Android devices and cannot be extracted as part of a standard mobile phone imaging. This is not the case with iPhone, where WhatsApp data could be captured in a readable format.

Some third-party applications store data within the cloud as opposed to on the user's device. Data from these applications cannot be extracted from a user's device during a standard collection. Applications that store data within the cloud may require separate standalone collections directly from the cloud services.

Collaboration Tools

Collection from collaboration tools like Slack and Teams requires navigating a collection of diverse sources, containing diverse content, and potentially, stored in diverse locations. Relevant communications may exist in public channels, private channels, direct messages, or group messages. It is not uncommon for an organization to have channels numbering in the thousands and messages numbering the millions. Moreover, each message may contain reactions, animations, links to videos, embedded content from third-party sources, and more.

Another challenge arises from the variety of licenses available. The type of license under which an organization uses Slack will dictate what options are available for preservation and export of relevant materials. For example, a free license for Slack caps how many messages can be preserved and exported, while paid licenses do not. Paid licenses also allow for more granular preservation options. In Teams as well, the Microsoft 365 license under which an organization uses Teams will determine what preservation and export tools are available.

An additional challenge arises from the diversity of places where relevant data may reside, which can complicate preservation and export. For example, different types of Teams data are stored in different places within the Microsoft 365 environment. Individual Teams content is stored in a user's mailbox, non-private channels content is stored in the group mailbox used for the team, and other types of content are stored in various SharePoint and OneDrive locations. In Slack or Teams, embedded content may be stored in third-party applications (e.g., Dropbox, YouTube) and just displayed dynamically based on the link that's actually in the message.

Because of these challenges and variations, as well as the additional challenges that arise during processing (e.g., expansion, format conversion, unitization), successful collection from these

[22]David Horrigan, e-Discovery Spoliation in Unusual Places: Preserve Your Pickup Truck, RELATIVITY BLOG, https://www.relativity.com/blog/e-discovery-spoliation-in-unusual-places-preserve-your-pickup-truck/ (Mar. 2, 2017).
[23]Katherine E. Vinez, The Admissibility of Data Collected from Wearable Devices, 4 Stetson J. Advoc. & L. 1 (2017), available at https://www2.stetson.edu/advocacy-journal/wp-content/uploads/2017/06/Vinez_-_Wearables.pdf

kinds of sources typically requires the assistance of an experienced collection expert, and it may require custom solutions.

Microsoft 365

Preservation in and export from Microsoft 365 presents the same challenges discussed above for Teams. It encompasses a wide range of sources and date types. It can contain enormous numbers of files and enormous volumes of data. Preservation and export options are dictated by license level, and they are complicated by the diverse array of places different types of user data is stored – both inside the Microsoft 365 environment and in third-party applications. Successful collection from these kinds of sources typically requires the assistance of an experienced collection expert (as well as the cooperation of the account holder, for individual accounts), and it may require custom solutions.

Social Media

For better or worse, social media is an influential, indispensable part of modern life. As it's permeated its way ever deeper into our professional and personal lives, its impact upon discovery has grown in parallel. In April 2019, the International Legal Technology Association published the results of its 2018 Litigation and Practice Support Survey,[24] revealing that 90% of responding professionals (overwhelmingly from law firms) had handled at least one case involving the collection and processing of social media data in the prior year, a 7% increase over the prior year.[25] Moreover, 19% reported handling more than 20 such cases, a 46% increase over the prior year.

Social media sources can pose technical challenges because they typically incorporate multiple forms and formats of media and communication together, creating a complex source of diverse ESI. They commonly allow sharing of photos and videos, status updates, public posts, private messages, live chats, video streams, and more. In addition to the material posted and uploaded by users, social media services also record extensive information[26] about each user's activities on the service, such as what content they've liked or shared, logs of when and how they've accessed the service, and sometimes more.

All of this material accumulates rapidly into large volumes because social media users access these services frequently and share hundreds of millions of new posts, messages, photos, and videos every day. Each individual social media account for each user can easily contain hundreds or thousands of pages of materials in a mishmash of formats. Facebook, for example, published a paper in 2021 on its transition to a new file system for its data centers[27] in which each cluster

[24]Cindy MacBean, *2018 Litigation and Practice Support Survey Results*, ILTA (Apr. 2019), available at http://epubs.iltanet.org/i/1108621-lps19/36?_ga=2.231156186.434461956.1629978821-1135214194.1629978821.
[25]ILTA's 2017 *Litigation and Practice Support Technology Survey Results*, ILTA (Apr. 2018), available at http://epubs.iltanet.org/i/973671-lps18/55?_ga=2.39038435.1141759458.1531162513-441756871.1531162513.
[26]*What categories of my Facebook data are available to me?*, FACEBOOK HELP CENTER, https://www.facebook.com/help/405183566203254?helpref=faq_content (2021).
[27]*Consolidating Facebook storage infrastructure with Tectonic file system*, FACEBOOK ENGINEERING, https://engineering.fb.com/2021/06/21/data-infrastructure/tectonic-file-system/ (June 21, 2021).

"scales to exabytes," up from "tens of petabytes" in their previous system.

There are three main options for the acquisition of social media materials for use in litigation:

- Printing out the material or capturing a screen image of it – this is fast and inexpensive, but it does not capture any native files or metadata. It may also create authentication and admission problems down the road.
- Using the self-service export tools provided by the social media platform – this, too, is fast and inexpensive, but it also may not provide native files or metadata. It often comes in a format that requires conversion using forensic tools, and not all parts of the content may be exported in a way that facilitates that conversion.
- Using specialized forensic collection software – this carries additional costs, but it can be essential for cases involving large quantities of social media materials, questions best resolved through the materials' metadata, or the potential for disputes over the authenticity and admissibility of the social media materials themselves. Escalating security and privacy measures, however, have begun to reduce how much these tools can do beyond the standard export function.

10.4 CONCLUSION

The challenges of identifying, preserving, and collecting relevant ESI continue to grow as old sources evolve, new sources emerge, and the behaviors of organizations and individuals adapt. The story of the past decade has been one of the long, slow march into the cloud, as organizations have transitioned to new software-as-a-service solutions and individuals have transitioned to new messaging and collaboration tools.

For practitioners, learning about these sources and the challenges they create is no longer optional, as the sources are becoming ubiquitous. The complexity and variability of these ESI sources, however, frequently makes these issues too complex to address without the assistance of relevant discovery or collection experts. Consulting with them early and often is the best way to make sure you know what is technically possible in each situation and how to proceed reliably and defensibly.

10.5 KEY TAKEAWAYS

There are seven key takeaways from this chapter to remember:

1. The story of the past decade has been one of the long, slow march into the cloud, as organizations have transitioned to new software-as-a-service solutions like Microsoft 365 and individuals have transitioned to new messaging and collaboration tools like Slack and Microsoft Teams.
2. Linked attachments (a.k.a., "modern attachments") are the result of a feature that inserts a link to the original document that a recipient can click to open it directly, instead of attaching a copy of the document to the message, which creates ambiguities around collection obligations and scope.
3. The proliferation of mobile device sources, social media sources, and collaboration tool sources has made message thread unitization a common question for eDiscovery, since these source types frequently include ongoing threads of back-and-forth messages that can span extended periods of time, and judgment must be exercised about how best to break them up into more usefully-sized chunks.
4. The increasingly-frequent presence of emojis creates special challenges for eDiscovery and litigation, both technical challenges caused by the volume and diversity of emojis, and challenges of interpretation caused by the inherently ambiguous and context-dependent meanings of the symbols.
5. Another change over the past decade has been the increase in availability and use of ephemeral messaging, which allows for the automatic deletion of sent messages after a set amount of time, and end-to-end encryption, which protects messages by ensuring it's impossible for anyone but the sender and recipient to read them, and both can cause significant preservation and collection challenges.
6. Beyond these general challenges, practitioners need to be aware of the source-specific challenges and limitations associated with specific sources, including mobile devices, OTT messaging apps, collaboration tools, Microsoft 365, and social media.
7. The growing complexity and variability of ESI sources means these issues are often too complex to address without the assistance of relevant discovery or collection experts; consulting with experts early and often is the best way to make sure you know what is technically possible and how best to proceed.

Chapter 11

Sampling Techniques for Litigation and Investigations

About this Chapter

In this chapter, we are going to review key sampling concepts and processes that are relevant to litigations and investigations, from winning a jellybean jar contest, to planning at the beginning of a project and checking completeness at the end, to testing your classifiers, both human and machine.

11.1 FINDING OUT HOW MANY RED HOTS ARE IN THE JELLYBEAN JAR

A candy store is running a contest. In the front window is a comically enormous jar of jelly beans, all different kinds and colors. Mixed in among them are a secret number of red hot cinnamon candies, similar in size and shape, all red. Whoever can guess closest to the true number of red hots mixed into the jar wins the prize. How do you guess? Do you try to count the red candies you can see, hoping they're all red hots, and then guess at how many you can't see? Do you try to count all the candies? Do you try to estimate volumes?

What if you were allowed to take one scoop of candies out of the enormous jar for closer examination, to determine exactly which ones in the scoop were red hots? Could you extrapolate from the scoop to the jar? How much better might your guess be then?

11.1.1 Sampling in eDiscovery

Despite years of discussion in the eDiscovery industry about the power and importance of sampling techniques – particularly in the context of technology- assisted review (TAR), many practitioners remain unfamiliar with what they can accomplish with them, and when, outside of TAR, they might do so. Beyond just being an essential part of TAR, however, there are opportunities across the phases of an eDiscovery project – whether for litigation or an investigation – to replace guesses based on anecdotal evidence with actual estimates based on formal sampling.

Courts have actually been encouraging parties to leverage sampling techniques in eDiscovery since before TAR existed, suggesting its use for the validation of search terms and document review processes:

- "Common sense dictates that **sampling and other quality assurance techniques must be employed** to meet requirements of completeness," *In re Seroquel Prods. Liab. Litig.*, 244 F.R.D. 650 (M.D. Fla. 2007)[1] [emphasis added]
- **"The implementation of the methodology selected should be tested for quality assurance;** and the party selecting the methodology **must be prepared to explain the rationale** for the method chosen to the court, **demonstrate that it is appropriate** for the task, and **show that it was properly implemented**," *Victor Stanley Inc. v. Creative Pipe Inc.*, 250 F.R.D. 251 (D. Md. 2008)[2] [emphasis added]

And they have continued encouraging its use for those purposes, even outside of TAR, to this day:

- "Just as it is used in TAR, **a random sample of the null set provides validation and quality assurance of the document production when performing key word searches**. Magistrate Judge Andrew Peck made this point nearly a decade ago. See William A. Gross Constr. Assocs., 256 F.R.D. at 135-6 (citing **Victor Stanley, Inc. v. Creative Pipe, Inc.**, 250 F.R.D. 251, 262 (D. Md. 2008)); In re Seroquel Products Liability Litig., 244 F.R.D. 650, 662 (M.D. Fla. 2007) (requiring quality assurance)." *City of Rockford v. Mallinckrodt ARD Inc.*, 326 F.R.D. 489 (N.D. Ill. Aug. 7, 2018)[3] [emphasis added]

[1] *In re Seroquel Prods. Liab. Litig.*, 244 F.R.D. 650 (M.D. Fla. 2007), available at https://casetext.com/case/in-re-seroquel-products-liability-litigation-16.
[2] *Victor Stanley Inc. v. Creative Pipe Inc.*, 250 F.R.D. 251 (D. Md. 2008), available at https://casetext.com/case/victor-stanley-inc-v-creative-pipe.
[3] *City of Rockford v. Mallinckrodt ARD Inc.*, 326 F.R.D. 489 (N.D. Ill. Aug. 7, 2018), available at https://casetext.com/case/city-of-rockford-v-mallinckrodt-ard-inc-1.

And, of course, the importance of sampling comes up again and again in discovery decisions and orders related to TAR use.

Industry publications, too, have taken repeated notice of the power and importance of sampling in eDiscovery. For example, sampling features prominently in The Sedona Conference's Commentary on Achieving Quality in the E-Discovery Process,[4] and the EDRM organization has released a commentary specifically on leveraging sampling in eDiscovery.[5]

11.1.2 Informal Approaches to Sampling

Many practitioners do engage in informal types of sampling already. As practitioners have done since the early days of discovery, it is common for a knowledgeable team member to test potential search terms and phrases by informally "poking around" in some of the results returned by them. The same thing goes for poking around in the materials collected from different sources or different custodians to determine the relative importance of different tranches of materials. The same also goes for quality control checks of document review efforts, with more senior attorneys poking around in the batches of documents reviewed by less-experienced attorneys to double-check their relevance or privilege determinations.

These informal approaches to sampling are inarguably valuable for gathering anecdotal evidence, making instinctual assessments, and learning about your materials or your efforts. Some information is always better than no information. But there are limits to what can be learned through these informal approaches and to how reliable such insights are.

11.1.3 Formal Approaches to Sampling

Formal approaches to sampling, on the other hand, facilitate more precise estimates with known reliability. It is these approaches that make sampling so valuable in TAR specifically and in eDiscovery generally. For example, formal sampling approaches can be used to generate:

- Reliable estimates of how many relevant documents are in a given tranche
- Reliable projections of the amount of redaction or privilege logging to do
- Reliable measurements of relevant materials missed by a given process
- Reliable reporting on the efficacy of a given search or other classifier

These measurements and many more can be taken using the same basic sampling techniques at various points in the discovery project lifecycle.

11.2 KEY SAMPLING CONCEPTS FOR WINNING THE CANDY CONTEST

In order to use sampling to estimate how many red hots are mixed into the jellybean jar, we need to under- stand some basic sampling concepts, including: sampling frame, prevalence, confidence level, and confidence interval, as well as how each affects required sample size. We also need to understand that whenever we refer to sampling here, we are referring to simple random sampling in which **any item within the sampling frame has an equal chance of being randomly selected** for inclusion in the sample.

[4]The Sedona Conference, *Commentary on Achieving Quality in the E-Discovery Process*, 15 SEDONA CONF. J. 265 (2014), available at https://thesedonaconference.org/publication/Commentary_on_Achieving_Quality_in_the_E-Discovery_Process.
[5]EDRM, *Statistical Sampling Applied to Electronic Discovery*, https://www.edrm.net/resources/project-guides/edrm-statistical-sampling-applied-to-electronic-discovery/ (Feb. 18, 2015).

11.2.1 Sampling Frame

Sampling frame refers to the set of materials from which a sample will be taken. In the context of our jellybean example, the sampling frame would be the full contents of the enormous jar of jellybeans and red hots. In the context of eDiscovery, your sampling frame will typically be the pool of materials available after any initial, objective culling has taken place (i.e., what's left for assessment and review after initial de-NISTing, deduplication, and date restriction during processing).

In addition to being your sample source, your sampling frame also affects the size of the samples you will need to take. As we will discuss below, sample size is primarily determined by how reliable and precise you want your results to be, but the size of your sampling frame also affects your needed sample size to some extent. As your sampling frame gets bigger, your sample size will also need to get bigger – but only up to a point. Beyond that point, the effect levels off, so the sample size needed for a frame of 100,000 items (e.g., jellybeans, documents, etc.) is roughly the same as the sample size needed for 1,000,000 of them, which is roughly the same as the sample size needed for 10,000,000 of them. Sampling frame size has the weakest effect on sample size.

11.2.2 Prevalence

Prevalence is how much of something there is within your sampling frame. For example, it could be how many red hots there are in your jellybean jar, or it could be how many relevant documents there are in your collected materials. It could also be how many documents are privileged, how many require redaction, or any other binary property you want to measure.

In the math underlying sampling, the prevalence of what you are seeking is also a factor that can have an effect on the required sample size for some purposes. When what you are doing sampling for is to estimate prevalence, however, you need to plug in an assumption for this value, and to be safe, you plug in the most conservative value (i.e., the one that results in the largest sample size). For prevalence, this is 50%, meaning that half the sampling frame is what you're looking for and half is not. Most sampling features in eDiscovery tools and online calculators will default to this value and may not even given you the option to change it.

11.2.3 Confidence Level

Confidence level is a measurement of how reliable your results are. It is expressed as a percentage out of 100, and most commonly, you will see discussion of 90%, 95%, or 99% confidence levels. What these numbers technically mean is that, if you reran the same sampling process 100 times in a row, you would expect to get similar results 90 times out of 100, or 95 times out of 100, or 99 times out of 100.

The higher you want your confidence level to be, the larger the sample size you will need to use to achieve it, and confidence level has a stronger effect on sample size than sampling frame size or prevalence does. For example, if you were taking a sample from a sampling frame of 100,000 items, and you wanted a margin of error of +/-2% (which we will discuss further below), here is how your required sample size would vary with your desired confidence level:

- For a confidence level of **90%**, a sample size of **1663**
- For a confidence level of **95%**, a sample size of **2345**
- For a confidence level of **99%**, a sample size of **3982**

11.2.4 Confidence Interval

Confidence interval is a measurement of how precise your results are. It is expressed as a percentage out of 100, and most commonly, you will see discussion of confidence intervals of 2%, 4%, and 10%. Even more commonly, you will see discussions refer to margin of error with references to +/-1%, +/-2%, and +/-5%. These margins of error are actually the equivalents of those confidence intervals. The latter is just framed in terms of plus or minus half the range, and the former is framed in terms of the full range.

The narrower you want your range of uncertainty to be, the larger the sample size you will need to use achieve it, and confidence interval (or margin of error) has the strongest effect on needed sample size. For example, if you were taking a sample from a sampling frame of 100,000 items, and you wanted a confidence level of 95%, here is how your required sample size would vary with your desired range of uncertainty:

- For a confidence interval of **10%**, a.k.a. a mar- gin of error of **+/-5%**, a sample size of **383**
- For a confidence interval of **4%**, a.k.a. a mar- gin of error of **+/-2%**, a sample size of **2345**
- For a confidence interval of **2%**, a.k.a. a mar- gin of error of **+/-1%**, a sample size of **8763**

11.3 RED HOTS, HOT DOCS, AND THE ONES THAT GOT AWAY

Now that we understand the necessary sampling concepts, let's apply those concepts to our candy contest and figure out how many red hots we think are in the jellybean jar. In order to do so, we will need to identify our sampling frame, select our desired confidence level, and select our desired confidence interval.

For this example, our sampling frame is all the candies in the enormous jellybean jar, which a sign indicates holds approximately 100,000 candies. For our confidence level, let's use 95%, which has been referenced in a variety of cases and articles as a potentially acceptable level of confidence, and for our confidence interval, let's use 4% – also known as a margin of error of +/-2%, which has also been widely discussed and used. (For example, 95% and +/-2% were the proposed values used in the plan in the da Silva Moore[6] case and in many other TAR cases.)

11.3.1 So, How Many Red Hots?

Now that we have our required values (**100,000, 95%, +/-2%**, and an assumed **50%** prevalence), we are ready to plug them into our sampling tool or calculator to find out how large our simple random sample will need to be. Most modern document review tools have some form of sampling tools built into them, but sampling calculators[7] are also readily available online and random document selections can be made in manual ways if needed

[6] *Moore v. Publicis Groupe*, 287 F.R.D. 182 (S.D.N.Y. 2012), available at https://docs.justia.com/cases/federal/district-courts/new-york/nysdce/1:2011cv01279/375665/96.
[7] *See e.g.* Raosoft, *Sample Size Calculator*, http://www.raosoft.com/samplesize.html (2004).

(e.g., by using the RAND function in Microsoft Excel). Plugging the values we've chosen into a sampling calculator reveals that we need a simple random sample of **2,345** pieces of candy to make our desired estimate, which is just **2.345%** of the total sampling frame.

Once a candy store employee has retrieved for us a randomly selected assortment of **2,345** pieces of candy from the jar, we can then review those sample candies up close to determine exactly which ones are cinnamon red hots. Let's say our review reveals there are 142 red hots among the **2,345** sample candies, or **6.1%**. We can now say – **with 95% confidence** – that the overall prevalence of red hots in the jar is **between 4.1% and 8.1%, or between 4,100 and 8,100 total red hots**.

If we were willing to review a larger sample of 8,763 candies, we could even narrow that range to between 5,100 and 7,100 total red hots.

11.3.2 So, How Many Hot Documents?

This same process can be employed in an eDiscovery project to make any number of useful estimations about a new collection of materials, including: **the prevalence of relevant materials, the relative prevalence of relevant materials in different sources, and the prevalence of materials requiring special review efforts (e.g., privilege logging, redactions, technical knowledge, etc.)**. These estimates can in turn be used to more accurately estimate your needed project resources, optimal project workflows, and likely project costs and durations. They can also be valuable in assessing the viability of a TAR solution or the need for additional objective culling. As projects progress, they can also provide a yardstick against which to measure progress and completeness.

It should also be noted that, when applying these techniques to eDiscovery, it is important to use the highest quality document review possible. While identifying cinnamon red hots is very straightforward, making legal or process determinations about documents can be quite nuanced, and the nature of sampling (extrapolating from a little to a lot) means that mistakes in classification during sampling will have amplified effects on the reliability of your estimates.

11.3.3 And, How Many Did We Miss?

One of the most common applications of prevalence estimation is in testing for completeness at the end of a TAR process or after the application of keyword searches. This is sometimes referred to as measuring **elusion**, i.e. the quantity of materials that eluded identification by the filtering and review process employed. For such estimations, the sampling frame is the pool of unreviewed materials eliminated before human review, by either the TAR software used, or by the keyword searches applied. The process is otherwise identical to the one described above.

There is no way to perfectly identify and produce **all** relevant electronic materials – and no legal requirement that you achieve such perfection, but there can be great value in being able to say with some certainty how little (or how much) has been missed. A reliable estimate can provide concrete evidence of the adequacy or inadequacy of a completed process, or a basis for arguing the proportionality or disproportionality of any additional discovery efforts.

11.4 TESTING CLASSIFIERS

11.4.1 What Is a Classifier?

Classifiers are mechanisms used to classify documents or other materials into discrete categories, such as those requiring review and those not requiring review, or relevant and non-relevant, or privileged and non-privileged. That mechanism might be a search using key words or phrases. It might be the decisions of an individual human reviewer or the aggregated decisions of an entire human review process. It might be the software-generated results of a technology- assisted review process. The binary classification decisions of any of these classifiers are testable in the same basic way. To start, we will focus on searches as the classifiers to be tested.

11.4.2 What Properties of a Search Classifier Do We Test?

When testing search classifiers, we are actually measuring two things about them: their **recall** and their **precision**, which correlate to their **efficacy** and their **efficiency**:

- **Recall is how much of the total stuff available to find the classifier actually found**, so higher recall (i.e., finding more) means greater efficacy, and lower recall (i.e., finding less) means lower efficacy
- **Precision is how much other, unwanted stuff the classifier included along with the stuff you actually wanted**, so higher precision (i.e., less junk) means higher efficiency, and lower precision (*i.e.*, more junk) means lower efficiency

Both recall and precision are expressed as **percentages out of 100**:

- For example, if there are **500** relevant documents somewhere in a dataset, and a search finds **250** of those documents, then that search has a recall of **50%** (i.e., 250/500)
- If the search returned **750** non-relevant documents along with the **250** relevant ones, that search would have a precision of **25%** (i.e., 250/1000)

There is also generally a tension between the two criteria. Optimizing a search to maximize recall beyond a certain point is likely to require lowering precision and accepting more junk, and optimizing a search to maximize precision beyond a certain point is likely to require accepting lower recall and more missed relevant materials. Deciding what balance between the two is reasonable and proportional is a fact-based determination specific to the needs and circumstances of each matter.

11.4.3 What Sample Is Needed to Test a Search Classifier?

In order to test a search classifier's recall and precision, you must already know the numbers of documents in the classifications you are testing. For example, to determine what percentage of the relevant material is found, you must know how much relevant material there is. Since it is not possible to know this about the full dataset without reviewing it all (which would defeat the purpose of developing good searches), classifiers must be tested against a control set drawn from the full dataset.

Much as we did for estimating prevalence, control sets are created by taking a simple random sample from the full dataset (after initial, objective culling) and manually reviewing and classifying the materials in that sample. Just as with estimating prevalence, it is important that the review performed on the control set be done carefully and by knowledgeable team members. In fact, in many cases you may be able to use the same set of documents you reviewed to estimate prevalence as a control set for testing classifiers.

Unlike estimating prevalence, however, figuring out the size of the sample needed for your control set is not so cut and dry. As we will discuss below, the reliability of the results you get when testing classifiers is related to how many potential things there were for the classifiers to find in the control set. For example, if you are testing searches designed to find relevant documents, the more relevant documents there are in your control set the more reliable your results will be.

This means that **datasets with low prevalence may require larger control sets to test classifiers than datasets with high prevalence**, depending on how reliable you need your results to be. The results of a prevalence estimation exercise can help you figure out how large of a control set you need (and whether your prevalence estimation set can just be repurposed for this exercise).

11.5 SHOW YOUR WORK: CONTINGENCY TABLES AND ERROR MARGINS

Once you have run a search classifier you are testing against your control set, you can calculate recall and precision for it by using contingency tables. Contingency tables (also sometimes referred to as cross-tabulations or cross-tabs) are simple tables used to break down the results of such a test into four categories: true positives, false positives, false negatives, and true negatives. These four categories are comparisons of the results of the search classifier to the prior results of your manual review of the control set:

1. **True positives** are documents that your search classifier returns as relevant results that your prior review of the control set also marked as relevant, *i.e.* **the right stuff**
2. **False positives** are documents that your search classifier returns as relevant results that your prior review of the control set had determined were not relevant, *i.e.* the **wrong stuff**
3. **False negatives** are documents that your search classifier does not return as relevant results that your prior review of the control set marked as relevant, *i.e.* **missed stuff**
4. **True negatives** are documents that your search classifier does not return as relevant results that your prior review of the control set had determined were not relevant, *i.e.* **actual junk stuff**

11.5.1 An Example Application

As we discussed above, sample sizes of a few thousand documents are common for taking prevalence measurements about large document collections. So, let's assume a hypothetical in which you have **a randomly selected set of 3,982 documents that you previously reviewed** to

take a strong measurement of prevalence (99% confidence level, with a margin of error of +/-2%) within your collection of 100,000 documents. Let's also assume that **your review of that random sample revealed 1,991 relevant documents**.

In addition to knowing prevalence within the overall collection (48-52% prevalence, with 99% confidence), you now have **a 3,982 document control set for testing search classifiers, containing 1,991 relevant documents for them to try to find**. The next step is running your search classifier against it and seeing how its classifications compare to those of your prior review. Let's assume your hypothetical search returns 1,810 total documents which break down into the four categories as follows:

As we can see on this contingency table, the 1,810 results from your hypothetical search included 1,267 documents that were also deemed relevant in your prior review, which are your true positives. It also

Once you have run a search classifier you are testing against your control set, you can calculate recall and precision for it by using contingency tables. Contingency tables (also sometimes referred to as cross-tabulations or cross-tabs) are simple tables used to break down the results of such a test into four categories: true positives, false positives, false negatives, and true negatives. These four categories are comparisons of the results of the search classifier to the prior results of your manual review of the control set:

1. **True positives** are documents that your search classifier returns as relevant results that your prior review of the control set also marked as relevant, *i.e.* **the right stuff**
2. **False positives** are documents that your search classifier returns as relevant results that your prior review of the control set had determined were not relevant, *i.e.* the **wrong stuff**
3. **False negatives** are documents that your search classifier does not return as relevant results that your prior review of the control set marked as relevant, *i.e.* **missed stuff**
4. **True negatives** are documents that your search classifier does not return as relevant results that your prior review of the control set had determined were not relevant, *i.e.* **actual junk stuff**

11.5.1 An Example Application

As we discussed above, sample sizes of a few thousand documents are common for taking prevalence measurements about large document collections. So, let's assume a hypothetical in which you have **a randomly selected set of 3,982 documents that you previously reviewed** to take a strong measurement of prevalence (99% confidence level, with a margin of error of +/-2%) within your collection of 100,000 documents. Let's also assume that **your review of that random sample revealed 1,991 relevant documents**.

	Deemed Relevant by Prior Review	Deemed Not Relevant by Prior Review
Returned by Search Classifier (i.e., Deemed Relevant)	1,267 (True Positives)	543 (False Positives)
Not Returned by Search Classifier (i.e., Deemed Not Relevant)	724 (False Negatives)	1448 (True Negatives)

In addition to knowing prevalence within the overall collection (48-52% prevalence, with 99% confidence), you now have **a 3,982 document control set for testing search classifiers, containing 1,991 relevant documents for them to try to find**. The next step is running your search classifier against it and seeing how its classifications compare to those of your prior review. Let's assume your hypothetical search returns 1,810 total documents which break down into the four categories as follows:

As we can see on this contingency table, the 1,810 results from your hypothetical search included 1,267 documents that were also deemed relevant in your prior review, which are your true positives. It also included 543 documents that were deemed not relevant in your prior review, which are your false positives. And, finally, we can see it missed 724 documents that were deemed relevant in your prior review, which are your false negatives.

You can use the results shown in this contingency table to easily estimate the recall and precision of the hypothetical search classifier you tested. As we discussed above, recall is the percentage of all the available relevant documents that were successfully identified by the search classifier being tested. So, in this example, your search identified 1,267 out of 1,991 relevant documents, **which gives you a recall of about 63.6%**. Also as discussed above, precision is the percentage of what the search classifier identified that was actually relevant. So, in this example, the search returned a total of 1,810 documents including 1,267 relevant documents, **which gives you a precision of 70%**.

Your hypothetical search, then, has high precision and good recall. The search could probably be revised to trade off some of that precision for higher recall or, possibly, to improve both numbers. Subsequent iterations of the search can be easily tested in the same way to measure the effect of your iterative changes.

11.5.2 How Reliable Are These Estimates?

We noted at the beginning of this hypothetical that your control set was a random sample of 3,982 documents that had been taken and reviewed previously to estimate prevalence in the full document collection with a confidence level of 99% and a margin of error of +/-2%. That same confidence level and margin of error, however, **do not carry over** to the estimates of recall and precision that you have made using the same documents. Because of how the math in question works, **your sample sizes for recall and precision are effectively smaller, which in turn makes your margins of error a little wider**.

The effective sample size for a recall estimation performed in this way is not the total number of documents in the control set, but rather **the number of relevant documents within it that are available to be found**. In this example, the search classifier is looking for the 1,991 relevant documents contained in the control set, which are effectively a random sample of 1,991 relevant documents from among all the relevant documents in the full document collection (a sampling frame you've already estimated to be about 50,000 documents).

The effective sample size for a precision estimation is also not the total number of documents in the control set, but rather **the number of documents identified by the search classifier**. In this example, the search classifier identified 1,810 documents, which are effectively a random

sample of 1,810 documents from among all the documents the search would return from the full document collection (a sampling frame you can estimate to be about 45,500 documents).

Some tools will provide you with these calculations automatically, but you can also plug these numbers into a sampling calculator[8] yourself to work backwards and see what margin of error would apply to your recall and precision measurements. In this example, your recall estimate would carry a margin of error of about +/-2.83% (at a confidence level of 99%), and your precision estimate would carry a margin of error of about +/-2.97% (also at a confidence level of 99%). Thus, you could be very confident that your tested search had a recall between 60.77% and 66.43% and a precision between 67.03% and 72.97%.

11.6 GRADING PAPERS: MEASURING HUMAN REVIEW

As we discussed above, a classifier can be a search, a TAR process, or other things – **including a human reviewer or a team of human reviewers**. Just as a search or a TAR tool is making a series of binary classification decisions, so too are your human reviewers, and the quality of those reviewers' decisions can be assessed in a similar manner to how you assessed the quality of a search classifier above. Depending on the scale of your review project, employing these assessment methods can be more efficient than a traditional multi-pass review approach, and in general, they are more precise and informative.

11.6.1 Human Classifiers and Control Sets

In this context, the reviewers doing the initial review work are the classifier being tested. **The control set is effectively generated on the fly by the more senior attorney performing quality control review**. Their classification decisions are the standard against which the initial reviewer's classification decisions can be judged. If an appropriate document tagging palette is employed (or if a sufficiently sophisticated review tool is being used), it is not hard to track and compare both sets of decisions to assess your human classifiers the same way we assessed searches.

In this context, however, we are not typically measuring the recall and precision of the human reviewers, although that could be done as well. **For human reviewers, it is more common to measure accuracy and error rate**. Accuracy is expressed as a percentage out of 100, and it represents t**he total number of correct classification decisions made by the initial reviewers**. Error rate is also expressed as a percentage out of 100, and it represents **the total number of incorrect classification decisions made**. Together, accuracy and error rate should add up to 100%.

	Deemed Relevant by the QC Reviewer	Deemed Not Relevant by the QC Reviewer
Deemed Relevant by the Initial Reviewer	70 (True Positives)	40 (False Positives)
Deemed Not Relevant by the Initial Reviewer	65 (False Negatives)	175 (True Negatives)

[8] *See e.g.* Raosoft, *Sample Size Calculator*, http://www.raosoft.com/samplesize.html (2004).

In terms of a contingency table, accuracy is derived from **the combination of all true positives and true negatives**, and error rate is derived from **the combination of all false positives and false negatives**.

11.6.2 An Example Application to Human Review

Let's look at an example of how this works. **Let's assume that you perform quality control review of a random sample of 350 of the 2,000 documents reviewed this week by a particular member of your initial review team**. After completing your classifications and comparing them to those of the initial reviewer, you get the following breakdown of results:

As with your search classifier, it is now straightforward to calculate an estimated accuracy and error rate for this reviewer's work this week. As noted above, accuracy is a combination of all the correct classification decisions, i.e. true positives + plus true negatives. So, in this example, your reviewer made 245 correct decisions out of 350 total decisions. **That gives you an accuracy of 70%**. As also noted above, error rate is combination of all the incorrect classification decisions, i.e. false positives + false negatives. So, in this example, your reviewer made 105 incorrect decisions out of 350 total decisions. That gives you an error rate of 30%.

There is also no reason that the same measurements could not be performed for more than one classification criteria based on the same quality control review (e.g., relevant and not relevant, privileged and not privileged, requiring redaction and not requiring redaction, etc.). **Any binary classifications for which you and your reviewers are making classification decisions can all be measured the same way**. The specific criteria measured and the specific results you get can then guide you in your ongoing reviewer training efforts, review oversight steps, and project staffing decisions.

11.6.3 How Reliable Are These Estimates?

As with testing search classifiers, you can work backwards from these results to determine how reliable these estimates of accuracy and error rate are, based on the size of the sampling frame (i.e., the total number of reviewed documents from which you pulled the sample) and the size of the sample you took. In this example, the sampling frame would be 2,000 and the sample size would be 350. Using those numbers, we find that your estimates of accuracy and error rate have a margin of error of +/- 4.76% at a confidence level of 95%. **Thus, you could be 95% confidant that the rest of the work from the reviewer in question was between 65.24% and 74.76% accurate**.

11.6.4 A Note about Lot Acceptance Sampling

Lot acceptance sampling is an approach to quality control that is employed in high-volume, quality-focused processes such as pharmaceutical production or military contract fulfillment. **In this approach, a maximum acceptable error rate is established, and each batch of completed materials is randomly sampled to check that batch's error rate at a predetermined level of reliability**. If the batch's error rate is below the acceptable maximum, the batch is accepted, and if the error rate is above the accept- able maximum, the batch is rejected.

Large-scale document review efforts have a lot in common with those other high-volume, quality-focused processes, and some particularly-large review projects have employed lot acceptance sampling in a similar way. Individual batches of documents reviewed by individual reviewers are the batches being accepted or rejected, and random samples are checked from each

completed one. Those with a sufficiently low error rate move on to the next phase of the review and production effort, those with too high of an error rate are rejected and re-reviewed (typically by someone other than the original reviewer). Error rates and batch rejections can be tracked by review- er, by team, by classification type, or by other useful properties to identify problem areas for process improvement or problem reviewers for retraining or replacement.

Many practitioners become uncomfortable at the idea of deliberately identifying an acceptable error rate, or even of actively measuring the error rate at all, but **avoiding knowledge of your errors does not prevent their existence**. It just prevents you from being able to address them or being prepared to defend them. After all, **the standards for discovery efforts are reasonableness and proportionality – not perfection**.[9]

11.7 KEY TAKEAWAYS

There are five key takeaways from this chapter to remember:

1. Formal sampling can replace intuitive assessments and assumptions with precise, reliable estimates, and judges have often ex- pressed a preference for argument and negotiation based on actual data and specific estimations rather than guesswork
2. Formal sampling has a variety of applications in litigation and investigations beyond just validation of technology-assisted review processes, including planning at the beginning of a project, checking completeness at the end, and testing your classifiers, both human and machine
3. Prevalence estimation can be accomplished by reviewing only a small percentage of large document collection, and it can be used to reliably estimate how many relevant documents are in a given tranche, the amount of redaction or privilege logging to do, the quantity of relevant materials missed by a given process, and more
4. Testing classifiers can also be accomplished by reviewing only a small percentage of a large document collection, and it can be used to iteratively improve your own searches, to evaluate those proposed by others, and to QC human document review
5. When using these sampling techniques, it is important to make sure you know how strong (confidence level) and how accurate (confidence interval/margin of error) your estimates need to be and will be, and consultation with an experienced expert is recommended

[9] "The second myth is the myth of a perfect response. The [respondent] is seeking a perfect response to his discovery request, but **our Rules do not require a perfect response. . . . Likewise, 'the Federal Rules of Civil Procedure do not require perfection**.' Like the Tax Court Rules, the Federal Rule of Civil Procedure 26(g) **only requires a party to make a 'reasonable inquiry' when making discovery responses,**" Dynamo Holdings Ltd. P'ship v. Comm'r of Internal Revenue, 2016 WL 4204067 (USTC 2016) [internal citation omitted; emphasis added].

Chapter 12

An Embarrassment of Riches: Analytic Tools and Techniques

About this Chapter

In this chapter, we will review key things practitioners need to know about the range of analytic tools and techniques available to them for ESI analysis and review. We will start by reviewing use cases and goals. Then, we will review the range of tools and techniques. Finally, we will review how to put it all together.

12.1 DIGGING FOR TREASURE

The tide of data never stops rising, and the types and sources of data never stop multiplying. Never have there been so many communication devices, apps, and services available. Never have there been so many ways to collaborate with others and generate electronically-stored information (ESI). Unfortunately, that also means there has never been more data that legal practitioners must somehow find a way to analyze and review. Finding a way that is efficient and effective requires understanding the range of tools and techniques available to you so you can pick the right tool for the right job.

12.1.1 Ethical Requirement

Beyond just being essential to the efficiency and efficacy of the effort, understanding how to go about sorting, filtering, and searching ESI is also a required part of attorneys' duty of technology competence for eDiscovery. In August 2012, the American Bar Association (ABA) implemented changes[1] to its Model Rules of Professional Conduct, including a change making the need to maintain technology competence explicit. Since then, that requirement or a variation on it has been implemented in forty states.[2] Understanding how to search effectively for the right data is a key part of fulfilling that requirement, and it was among the nine core skills originally identified by California's Formal Opinion No. 2015-193.[3]

12.2 USE CASES AND GOALS

There are three main use cases for these analytic tools and techniques and a variety of goals or priorities you may wish to pursue.

12.2.1 Early Case Assessment

The term early case assessment (ECA) originally referred to reducing uncertainty about the risks and costs associated with a new legal matter by quickly making a preliminary assessment of the evidence, facts, and law to inform decisions about how to proceed. Today, it also encompasses early data assessment, which is focused on evaluating the composition and completeness of collected ESI, and review preparation, which includes tasks like testing and refining searches and filters, evaluating potential workflows, and estimating needed resources. The intersection of these three connected-but-distinct activities makes the ECA phase of an eDiscovery effort one in which many different analytic tools and techniques can be useful.

12.2.2 Document Review

After ECA, document review itself can also benefit extensively from the application of analytic tools and techniques. As noted above, such tools and techniques can be used to plan and estimate, but they are also essential for organizing and prioritizing materials for review. Moreover, full application of a technology- assisted review or continuous active learning workflow can dramatically improve the efficiency and efficacy of a document review effort.

[1]Debra Cassens Weiss, *Lawyers Have Duty to Stay Current on Technology's Risks and Benefits, New Model Ethics Comment Says*, ABA JOURNAL, http://www.abajournal.com/news/article/lawyers_have_duty_to_stay_current_on_ technologys_risks_and_benefits/ (Aug. 6, 2012).
[2]Robert Ambrogi, *Tech Competence*, LAWSITES, https://www.lawsitesblog.com/tech-competence (last visited July 2, 2021).
[3]The State Bar of California Standing Committee On Professional Responsibility and Conduct, *Formal Opinion No. 2015-193* (June 30, 2015), available at https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL 2015-193 %5B11- 0004%5D (06-30-15) - FINAL.pdf.

12.2.3 Investigations

In investigations, whether internal or agency-initiated, organizations face discovery needs similar to those they face in litigation. ESI that is both voluminous in scale and diverse in type and source must be analyzed and reviewed in an efficient and effective way. Beyond that, investigations add the additional challenges of tighter timelines, of limited negotiability (for agency investigations), and potentially, deliberate obfuscation by bad actors. These challenges only increase the importance of leveraging the right tools to facilitate efficient analysis and nuanced review.

12.2.4 Goals

Depending on the use case and the specific situations, you may wish to pursue any or all of these goals:

- Assessment of a case on its merits to determine risk/how to proceed
- Rapid identification of hot documents for an internal or agency investigation
- Assessment of a collection's completeness (both within and across custodians)
- Estimation of volumes, prevalence, etc. for review and production planning
- Testing of sorting, filtering, and searching to identify relevant materials
- Efficiently conducting review (organizing, prioritizing, reduction through TAR/CAL)

Thankfully, in most document review platforms, practitioners have a powerful set of tools and techniques at their disposal for pursuing these myriad goals. The specific bells and whistles vary, but generally, there will be some kind of random sampling tools, searching and filtering tools, structured analytics tools, conceptual analytic tools, and technology- assisted review workflows. Beyond those options, there are other new and developing tools that may be available too. Let's discuss each of these tools and techniques.

12.3 SAMPLING TOOLS AND TECHNIQUES

One of the most powerful tools in your toolkit is sampling. There are a lot of ways to find materials you expect to be in a collection of ESI, but sampling is a terrific way to also find materials you didn't know to look for: the unknown unknowns. For our purposes, sampling comes in two flavors: judgmental sampling and formal sampling.

Judgmental sampling is the informal process of looking at some randomly selected materials to get an anecdotal sense of what they contain, whether that's sampling from a particular source, from a particular search's results, or from a particular time period. You're not reviewing a particular number of documents or taking a defined measurement with a particular strength; you're getting

an impression and making an intuitive assessment.

Formal sampling is just the opposite: you are reviewing a specified number of randomly-selected documents with the goal of taking a defined measurement with a particular strength. Typically, that measurement is either of how much of a particular thing there is within a collection (i.e., estimating prevalence) or of how effective a particular search is (i.e., testing classifiers).

- **Estimating Prevalence** - Estimating prevalence is the process of reviewing a simple random sample of a given collection of materials to estimate how much of a given kind of thing is present. You might estimate the prevalence of relevant materials, of privileged materials, or of materials requiring redaction or other special steps. The size of the sample you need is dictated primarily by how precise you want your estimate to be (i.e., margin of error), and how certain about it you want to be (i.e., confidence level), and to a lesser extent, by how large your collection of materials is (i.e., sampling frame). Most often you will be dealing with sample sizes of a few thousand (e.g., a sample of 2,345 for a confidence level of 95% and a margin of error of +/-2% in a collection of 100,000 documents).
- **Testing Classifiers** - Testing classifiers is the process of seeing how effective and efficient a particular classifier – typically a search of some kind – actually is. Using this technique, you can estimate how much of what you're seeking a given search is likely to return (i.e., recall) and how much irrelevant material is likely to get returned with it (i.e., precision). These measurements are taken by running the searches against a control set, which is made be pre-reviewing and coding a sufficiently-large random sample. Comparing the search results to the already-completed coding allows for the iterative refinement of searches to increase their recall and precision before they are applied to the full collection.

12.4 SEARCH AND FILTERING TOOLS

After sampling, the next major category of tools and techniques available is search and filtering, including keyword and phrase searching, Boolean searching, fuzzy searching, conceptual searching, and more.

12.4.1 Searching

Searching, both on the internet and among our own emails, messages, and files, has become an inescapable part of everyday life. Almost all of this searching, like the searching you do in eDiscovery, is powered by some form of indexing. In the eDiscovery context, indexing is typically performed during the processing phase of the project.

Indexing is the process of creating the enormous databases that are used to power search features. Most common are inverted indices, which essentially make it possible to look up documents by the words within them. Inverted indices are like more elaborate versions of the indices you find in the backs of books. Decisions during processing about how indices should be generated and what common words (e.g., articles, prepositions) they should skip affect the completeness of search results you get. Searches can only find what indices show.

More sophisticated indices are created to power features like concept searching, concept clustering, and technology-assisted review, which we will discuss further below in our section on

advanced analytic tools and techniques. The types of indices that are prepared and the specific features your software offers for working with them will dictate what types of searching are available to you.

- **Keyword and Phrase Searching** - Exactly as it says on the tin, keyword and phrase searching lets you search for a key word, for a phrase, or for lists of both at once. Just as with the basic internet searching we all use, if one of the desired keywords or phrases is present, the document will be returned. One key area of variation from tool to tool is whether wildcard characters can be used to find variations on words and, if so, how they can be used.
- **Boolean Searching** - Boolean search is the next step up in sophistication from basic keyword searching. It allows the use of operators such as "and," "or," and "not." These operators allow for the searcher to define specific relationships between key words and phrases to achieve higher quality results (i.e., improved recall and precision). Other operators may be available, including proximity operators (i.e., to find a particular word appearing within a certain number of words of another particular word). The range of specific operators available varies with the tools being used, as can their precise operation. Thus, it is important to understand the tools you are actually using to be sure you are searching the way you intend.
- **Fuzzy Searching** - Fuzzy searching (also sometimes referred to as approximate string matching or stemming) is another extension of basic keyword searching that may be available to you. Fuzzy searching allows a search to return variations on a word rather than just the precise word you searched (e.g., finding both invite and invitation). How much variation is allowed is typically an adjustable setting.

- **Conceptual Searching** - As noted above, conceptual searching is powered by different types of indices than traditional searching. Conceptual searching uses these indices to try to return results based on related ideas and topics rather than just based on whether the same specific words and phrases are used.
- **Other Tools and Features** - In addition to these core search functions, most review tools also offer a range of reporting and administration tools (e.g., saved searches, search history, etc.) to assist you in brainstorming, testing, and iteratively improving searches to meet your information needs. Many tools now also offer some form of word cloud or topical heat map feature to facilitate visual review of the most used words or phrases in your materials.

12.4.2 Filtering

In addition to your searching options, most platforms also offer you a range of options for sorting and filtering by specific properties of documents to help you surface what matters and prioritize what matters most. Most often this is based on a combination of metadata values extracted from the documents, such as file type and date, and custom-created metadata values, such as domain name or custodian.

Often, these types of sorting and filtering capabilities are now tied to visualization tools that let you see the distribution of materials (and any gaps in them) at a glance and that allow you to adjust a range of value limits to see how they narrow or expand your results. For example, many tools now offer communication maps that can show which people are communicating with each other, how often they are doing so, and other useful details.

12.5 STRUCTURAL ANALYTIC TOOLS

After sampling tools and searching and filtering tools, the next major type of tools available to aid your analysis and review are structural analytic tools that facilitate email threading, duplicate identification, repeated content identification, and textual near- duplicate identification.

12.5.1 Threading

Despite the rise of mobile and social sources, collaboration tools, and other alternative communication channels, email still remains a major component of most ESI collections for eDiscovery, and it tends to be voluminous. A single gigabyte of email can easily contain 5,000 to 10,000 discrete email messages, plus their attachments and embedded images and objects. Thankfully, email ESI also typically contains a significant amount of repetition and overlap that can be skipped.

For example, if you collect email from two custodians, you will have multiple copies of the email messages sent between them – a sender copy and a recipient copy for each one. Moreover, if they are engaged in a thread of replies to each other, the emails in such a thread may contain the preceding emails within themselves as quoted text, and the last one in the thread may contain the full text of the whole thread within itself. Such emails are sometimes referred to as inclusive emails, as are any standalone or offshoot emails that contain unique content or attachments. Email threading tools typically offer some version of two functions to users: conversation threading and inclusive email identification. Additionally, as noted above, many now offer visualization features as an alternative way to explore the email threads and inclusive emails identified by the system.

- **Conversation Threading** - Conversation threading is a process in which emails are analyzed and automatically organized into thread groups, arranged chronologically. This analysis looks at existing conversation IDs, if available, and a range of email header fields and other document properties to match up replies in sequence. Such organization makes it possible to quickly identify related materials, speeding up investigation, and to quickly see the context surrounding a particular message, improving understanding. Additionally, presenting emails to reviewers as organized threads speeds up later document review.
- **Inclusive Email Identification** - Inclusive email identification is a process in which textual analysis is used to identify inclusive emails, i.e. those that contain a full thread within themselves or that otherwise contain unique text or attachments. Identifying the inclusive emails allows you to more quickly get the full picture, speeding up investigation, and when used as a filter, it can dramatically reduce the number of emails requiring later document review.

12.5.2 Duplicates

The operation of computer systems can produce a lot of duplicate files (including duplicate emails, as noted above). Although duplicates may need to be tracked and reported on in certain

circumstances, they do not need to be examined. Such duplicate files are identified using a technique called hashing.

Hashing is a technique by which sufficiently unique "fingerprints" can be generated for files. Hash functions are mathematical processes that take irregular-length inputs (e.g., the data in a particular file), and use them to generate fixed-length outputs (e.g., a string of 32 numbers and letters). Hashing for duplicate identification is accomplished using a cryptographic hash function (e.g., MD5 or SHA-1), which is well- suited to matching unique inputs to particular outputs. Identical files produce identical hash values, and hash values can be easily compared by software to automatically identify matches across even a large collection of ESI.

- **Duplicate Identification** - Typically, collected ESI is hashed and deduplicated during or prior to the ECA phase of the project. But, because other rules (e.g., family group preservation) may override deduplication, some duplicates may remain. For example, if the same spreadsheet was attached to two different emails, neither copy would be removed. Most platforms provide features for identifying and managing such duplicates within your loaded collection of materials.
- **Repeated Content Identification** - In addition to identifying fully-duplicated documents, many platforms also offer some form of repeated content identification. Such features are designed to automatically identify frequently-repeated blocks of text (e.g., email signature blocks, automatic confidentiality warnings, etc.) so that they can be filtered out of search results (reducing false positives, particularly for privilege searches) and omitted from the creation of semantic/conceptual indices (improving the effectiveness of the semantic/conceptual analytic tools we will discuss below).

12.5.3 Near-Duplicates

In addition to true duplicates, it is common for collections of ESI to contain large numbers of near- duplicates. Near-duplicates are documents that are substantially similar to each other, but not truly identical (and therefore not removed during deduplication). There are two main types of near-duplicates that occur:

1. Superficially-identical documents that only vary in some metadata property, typically arising from their different sources or collection methods
2. Documents with some actual variation in con- tent, like successive drafts of a contract

Finding the former reduces the number of documents to consider (and later review), while ensuring consistent treatment across duplicates. Finding the latter can provide valuable context to the development of key documents over time.

- Near-Duplicate Identification - Textual near-duplicate identification is somewhat more complicated behind the scenes than true-duplicate identification. Rather than comparing whole documents as single, abstracted values, the full textual content of the documents must be broken down into smaller pieces (sometimes called shingles). These small pieces can then be hashed and the sequences of those pieces compared across documents. If a sufficient number of pieces match, in the right order, the documents will be treated as near-duplicates. Typically, the threshold of similarity at which the system treats two documents as near-duplicates can be customized.

12.6 CONCEPTUAL ANALYTIC TOOLS

The next major tools available to aid your analysis and review are advanced analytic tools, powered by conceptual indexing and other advanced mathematical analyses, including concept searching, concept clustering, and categorization.

12.6.1 Conceptual Indexing

As we noted briefly above, there are more sophisticated types of indices than the traditional inverted indices used to power basic search functions. These conceptual indices (sometimes called semantic indices) analyze the available materials in a different way to power different kinds of features. Whether created by latent semantic analysis, probabilistic latent semantic analysis, support vector machines, or another related mathematical approach, these indices are designed to go beyond just listing all of the words in a document to reveal their conceptual content.

This analysis is accomplished mainly by analyzing the co-occurrence of unique terms across the collection of documents (e.g., how often does the term "fire" appear with the term "employee" and how often does it appear with the term "extinguisher"). This analysis of co-occurrences is used to create an n-dimensional map (like a traditional map of Cartesian coordinates, but with many more dimensions than just x, y, and z). The more frequently unique terms co-occur together, then the stronger the relationship between them, and the more co-occurring terms in two documents, then the closer to each other they will appear on the map. Dense clusters of such documents suggest key topical areas in the document collection (e.g., employee termination discussions in one area of the map and fire safety discussions in another).

12.6.2 Features Powered by Conceptual Indexing

Conceptual indices are used to power a variety of branded features in different eDiscovery software platforms, but regardless of name or variation, there are three key functions that are generally available:

- **Concept Searching** - Searching against a conceptual index does not require an exact match in the way that searching against an inverted index does. Instead, the terms or phrases you search are mapped onto the existing index and documents that are close enough to those search terms on the map will be returned as results – even if none of the exact terms you searched appear. Some concept searching features are referred to as natural language search features, and some also offer an option to search for more documents like a given example document, which may be a real sample or a synthetic one, created for the purpose.

  Another advantage of searches against these indices is that they can reveal more than just a binary, yes-or-no result. Because of the nuanced multidimensionality of the index, you can get results scored on how responsive or not responsive they are to your search (i.e., how close or far away the result was on the map).
- **Concept Clustering** - Concept clustering is an automated, unsupervised process in which software analyzes the conceptual index that has been created. Rather than looking for the closest matches to a user-provided search, the software looks for the densest clusters of related materials it has identified and groups those results together into clusters defined by their most frequently occurring terms. How dense a cluster must be to qualify is typically a customizable property. Those clusters can then provide an alternative way to explore a collection of documents, to learn about

the scope of topics and range of materials it contains, and to identify areas for further exploration.

- **Categorization** - Categorization is akin to a hybrid between concept searching and concept clustering. It is a process in which a user selects a set of example documents to define a cluster for the software, and then the software attempts to find all the other documents that should go in that cluster with the examples provided. This is the functionality that powers some kinds of technology-assisted review.

12.7 TECHNOLOGY-ASSISTED REVIEW WORKFLOWS

Technology-assisted review is used to refer to a family of workflows that leverage categorization (or similar functions), in combination with sampling, to achieve a reliable document review process that requires significantly fewer hours of manual, human review than traditional all-manual approaches. Since its initial rise to prominence in 2011, the available array of TAR tools has expanded and evolved, and eDiscovery service providers have continued to develop new workflows to leverage them in useful ways for the diverse range of projects their clients face.

Although full deployment of a TAR workflow is typically part of the review phase of an eDiscovery project, these workflows – or limited versions of them – may also be leveraged to explore a collection during ECA, to organize and prioritize it for a more traditional review process, or to create a yardstick against which to measure a more traditional review process.

TAR approaches come in two major varieties, which we will refer to as TAR 1.0 and TAR 2.0:

- **TAR 1.0 – Predictive Coding** - TAR 1.0 refers to the initial, categorization-based workflows offered in eDiscovery – many of which were, and are, referred to as predictive coding. Broadly speaking, these workflows involve leveraging a sampling process to create a training set or seed set (i.e., a user-defined cluster or clusters), which the chosen software than uses to find other similar documents. These results are then reviewed and coded, and that coding is used to improve the software's results. This training cycle is iterated multiple times until an acceptable quality of results is achieved. The effectiveness of the whole process is measured using either a previously prepared control set or an additional random sampling effort.
- **TAR 2.0 – Continuous Active Learning** - TAR 2.0 refers to more recent workflows developed to leverage new tools based on different mathematical approaches. Rather than being based on identifying the similarities in a large, prepared training set like categorization and TAR 1.0, these workflows are characterized by continuous active learning that updates relevance scoring and prioritization for all documents dynamically as each additional document is coded by a reviewer.

  This is accomplished by focusing on a single, binary classification (i.e., relevant to topic X and not relevant to topic X) and analyzing the differences in language

between succes- sive, single example documents to identify the hyperplane that best divides the relevant examples from the non-relevant examples

on a multidimensional map. Each additional example the software analyzes and maps can lead the software to identify a more efficient hyperplane between the two groups, improving its classifications.

These workflows emphasize speed over struc-ture, and so, they work best in situations where there is a clear, binary classification decision to make and where family groups and other contextual factors are less important than overall speed.

12.8 NEW AND DEVELOPING TOOLS

The above tools and techniques are all well-established and widely-available. Software and service providers, however, are continually innovating new tools and techniques to address new and developing challenges. Some that you should be aware of are tools for PII analytics, entity extraction, image analysis, and generative AI.

12.8.1 PII Analytics

Personally identifiable information (PII) is defined by the United States Department of Labor as:

> Any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means.[4]

Common examples of this kind of information include: name, social security number, passport number, driver's license number, taxpayer identification number, patient identification number, financial account number, credit card number, VIN number, title number, street address, email address, telephone number, and biometric data. As more privacy laws and regulations are being promulgated in more jurisdictions, protection of PII is becoming ever more important, and PII frequently appears in the kinds of ESI collected for investigations and litigation.

PII analytic tools are designed to automatically identify certain kinds of PII wherever it appears in your collected ESI. There are two general ways the tools can work. They can operate based on pattern matching (e.g., finding text strings formatted like social security numbers), or they can operate based on algorithmic or AI analysis (i.e., context aware; can identify more than just text string formatting). The more sophisticated tools can provide additional help, such as finding more types of PII in more types of data, finding custom defined types of PII, automatically associating identified PII with identified custodians or entities, and providing the basis for programmatic redaction of specified data strings.

12.8.2 Entity Extraction

Like PII identification, analytics tools can also identify and extract a broad array of entities. Entities often include people's names, addresses, email addresses, phone numbers, as well as locations, organization names, and product names. Entity extraction applications, which are also known as Named Entity Recognizers (NERs), apply artificial intelligence to classify various types of entities, such as a location, a person, or an organization.

For example, in the sentence "Fashion designer Ralph Lauren founded Ralph Lauren in 1967"

[4]"Guidance on the Protection of Personal Identifiable Information," U.S. Department of Labor, available at https://www.dol.gov/general/ppii.

the name "Ralph Lauren" is used to refer to both a corporation and a person. To differentiate these two uses, entity extraction applications analyze the context of the name, including the surrounding words and grammatical structure. In this case, the word "founded" enables the application to determine that Ralph Lauren was both referred to as person ("Ralph Lauren founded") and referred to as an organization ("founded Ralph Lauren").

12.8.3 Image Analysis

Image analysis tools enable users to find objects of interest within images or videos. Searching images for objects of interest (e.g., a person, a car, a passport photo, inappropriate content) often occurs when working with mobile phone data or security video (e.g., video of employees enters a building). Image analysis applications attempt to identify objects by examining the image's pixels. When a group of pixels form a pattern, the application can then isolate distinct objects within the image. For example, when analyzing images of people entering a building, the application uses the image's pixels to identify a person versus a delivery cart or a person's face and their purse or bag.

Image analysis applications do this by training on a very large library of objects that appear in our everyday environment (e.g., people, animals, trees, cars, etc.). This training allows the application to classify the objects that are detected in the image (e.g., a delivery cart, a person's face, a bag, or purse). When a user submits a search, such as an employee's face, the application can then compare the employee's face against all other faces that were extracted from the video.

12.8.4 Generative AI

Recent advances in AI have led to the creation of generative AI applications that can generate new content based on their understanding of existing content. Generative AI can respond to natural language queries, draft natural language responses, generate images, generate audio, and generate video, and numerous other generative applications are still in development or testing. The most widely known of these tools so far is ChatGPT, which attempts to understand natural language queries and draft natural language responses. It accomplishes this using something called a Large Language Model (LLM).

Large language models are tools trained on enormous collections of text – often, text scraped from the Internet (e.g., Wikipedia, Reddit, news articles, blogs, case law databases, etc.). LLMs allow generative AI applications to understand grammar and sentence structure, enabling them to predict the best next word in a sentence with a very high degree of accuracy. For a very simple example, a user could ask it to complete the phrase "peanut butter and," and based on its training across Internet data, the application would complete the phrase with "jelly."

In the context of discovery and investigations, tools based on generative AI may soon provide a way to explore your collected ESI for relevant information and materials using natural language queries, or an "assistant" to aid in deposition or motion preparations. Expanding from the simple example above, imagine submitting a more complex request, such as: "Generate a draft legal complaint for an Illinois state court by John Doe against Jane Doe for property damage to a fence arising out of a car accident on May 1, 2023, at 1060 W. Addison St. Chicago, IL." A properly trained LLM could make fulfillment of such a request possible.

Generative AI applications also have limitations, including providing inaccurate or biased results. LLMs can only be as good as the data sets on which they are trained, and data scraped

from across the Internet contains all the inaccuracies and biases that are common online. Also, because LLMs simulate language rather than cataloging information, generative AI applications may also be prone to "hallucination" of information that sounds plausible but is not real. For example, one attorney using ChatGPT was provided with fake case citations, which they used, resulting in sanctions.[5] Due to these limitations, generative AI applications should still be used with caution during their nascency, and users should always validate and QC responses.

12.9 PUTTING IT ALL TOGETHER

Our survey of analytic tools and techniques has revealed a wide range of available options, each with different strengths and use cases, but achieving effective analysis and review is not a question of applying as many of these tools and techniques as you can. Rather, it is a question of selecting the right ones to best serve your current goal – whether that's traditional ECA, data assessment, investigation, or review preparation – and then building on those steps in a rational way to eventually achieve all your goals for the project.

Much like a telescope or microscope, the best result does not come from lining up as many lenses as possible. You must align the right ones on the right order to bring what you seek into sharp focus. Let's review some examples of how to think through these decisions.

12.9.1 For Pursuing Traditional ECA

When your top priority is pursuing traditional ECA, the first question to ask yourself is how much knowledge you have of what you expect to find. If you know a lot about what you're looking for in your ESI (e.g., from thorough custodian interviews, from overlap with prior legal matters, etc.), you may be able to jump right to searching for it. If you don't know a lot about the materials you're seeking, which is more common, you will want to start with one or more of the tools and techniques best suited to revealing unknown unknowns:

- Formal random sampling to estimate prevalence, which lets you see a cross-section of everything you have and some of all the different terms and phrases your custodians use to help you better plan your next ECA steps
- Visualization tools (e.g., communication maps, word clouds, etc.), which can reveal patterns of communication and behavior and assist with completing the picture of what happened in other ways
- Conceptual indexing features, which let you use concept searching to find relevant materials without knowing the best search terms, concept clustering to explore a cross- section of topics, and categorization to use a few relevant examples to find more

[5]Jane Wester, "Judge Imposes $5K Fine on Lawyers Who Submitted ChatGPT-Generated Fake Case Citations," N.Y.L.J. (June 22, 2023), available at https://www.law.com/newyorklawjournal/2023/06/22/judge-imposes-5k-fine-on-lawyers-who- submitted-chatgpt-generated-fake-case-citations/.

- Continuous active learning (TAR 2.0) workflows, which can rapidly surface relevant materials in certain circumstances

Once you start to get a handle on what you are really seeking (or if you already knew), you can transition from these initial, exploratory efforts to more targeted search and filtering efforts, which can quickly find relevant materials and hot documents. And, as you find relevant materials to review, thread and duplicate management tools can be used to find related materials to review for context as needed (e.g., related emails, alternate drafts, etc.).

12.9.2 For Assessing the Completeness of Your ESI

If your top priority is assessing your collected ESI, finding individual documents and facts is less important than ensuring a sufficiently complete collection has taken place and that any filtering applied during processing has not been excessive. In such situations, your focus should be on tools and techniques that help you see the big picture of your ESI collection and reveal the gaps within it:

- Metadata filtering and visualization tools, which help you assess the completeness of your collection by revealing ranges of values and gaps in those ranges, as well as potentially revealing important date ranges and sources, the connections between custodians, and more
- Concept clustering, which can provide a valuable overview of the content types and topics within your materials, including revealing an absence of things you expected or the presence of things you don't need
- Visualization tools (e.g., communication maps, word clouds, etc.), which can reveal collection gaps, including missing date ranges, missing custodians, and more
- Thread and duplicate management tools, which can provide another way to map conversation threads to reveal gaps requiring further collection, or which can reveal the presence excessive near-duplicates suggesting a collection or processing issue

Formal random sampling can also be useful, particularly if there are disputes over the appropriate scope of preservation and collection that need to be resolved. Sampling to estimate prevalence can be used to apply relative value determinations to different sources and tranches and to estimate costs and benefits associated with specific proposed work.

For Pursuing Review Preparation

When your top priority is review planning and preparation, you are concerned with learning about what happened, but only insofar as that informs what must be reviewed later and how it should be prioritized. And, you are concerned with understanding the properties and the big picture of the ESI you've collected, but only insofar as that informs what tools and techniques for culling you should choose and what review methodologies are likely to be effective. All of the tools and techniques discussed so far can be leveraged to assist in this effort:

- Formal random sampling to estimate prevalence, which allows you to accurately estimate what you have, to evaluate the suitability of potential review workflows (including assessing the viability of a TAR or CAL solution or the need for additional objective culling), and to create a yardstick against which to measure future review work
- Formal random sampling to test classifiers, which allows you to iteratively

improve any searches you plan to apply for culling, to ensure that they minimize unnecessary downstream review work and that they avoid missing any important materials

- Searching and metadata filtering, which can both be leveraged to eliminate as much of the chaff as possible without losing an unreasonable amount of the wheat, thereby reducing all downstream review and production costs
- Thread and duplicate management tools, which can dramatically speed up later review work, both by eliminating materials not requiring review and by providing superior organization to what remains
- Semantic indexing features, which can let you use concept clustering to help organize and prioritize subsequent review activity or let you leverage TAR workflows

12.10 KEY TAKEAWAYS

There are nine key takeaways from this chapter to remember:

1. There are a variety of use cases in which analytic tools and techniques will be valuable, and a variety of goals you may pursue with them.
2. To pursue these goals, document review platforms include a range of useful features, including: random sampling tools; searching and filtering tools; structured analytics tools; conceptual analytics tools; technology-assisted review workflows; and potentially, PII tools, entity extraction tools, image analysis tools, and generative AI tools.
3. Sampling tools and techniques – particularly formal sampling – are good at revealing unknown unknowns, at giving an accurate overview of your materials, and at providing a reliable basis for improving searches and planning subsequent steps.
4. Search and filtering tools, including newer visualization tools like communication maps, are good at finding specific materials, at identifying gaps in your collection, and at eliminating irrelevant materials prior to review.
5. Structured analytics tools for managing threads and duplicates are good at placing documents in context, at prioritizing and organizing materials for review, and at avoiding duplicative review.
6. Conceptual analytics tools are good at exploring a collection of materials without foreknowledge of the contents and key terms and rapidly surfacing relevant materials.
7. Technology-assisted review workflows are good at prioritizing and organizing materials for review, reducing the volume of materials to be reviewed, and ensuring review quality.
8. Sometimes you may be able to leverage other new and developing analytic tools, such as PII tools, entity extraction tools, image analysis tools, and generative AI tools.
9. Successfully achieving your goals requires leveraging the right combination of these tools and techniques based on what you're trying to find out and how much you already know.

Unit 3

Advanced eDiscovery

Chapter 13 - Everything in Moderation: Proportionality in Discovery

The 2015 amendments to the FRCP brought proportionality front and center in an attempt to combat the runaway cost and scale of eDiscovery. Courts and parties have since increased their focus on proportionality as a discovery limit. This chapter discusses proportionality, the key factors courts consider, and other issues.

Chapter 14 - When the Bough Breaks: Spoliation in eDiscovery

Enormous volumes and diverse types of ESI may be relevant to a case, and this can make identification and preservation challenging. It is almost inevitable that some ESI will slip through the cracks. When that happens, FRCP 37(e) provides the framework for assessing the loss and its consequences.

Chapter 15 - An Ounce of Prevention: Fundamentals of Data Protection

For legal practitioners, a commitment to data protection is a matter of ethical duty, legal and regulatory compliance, and fundamental responsibility to clients and stakeholders. Fulfilling that commitment requires consideration of security compliance frameworks, role-based access control, cloud storage vs. on-premises storage, and data encryption.

Chapter 16 - Cross-Border Discovery: A Guide to Practical Challenges for US Counsel

As the world's economies continue to reach across borders, US counsel representing companies of all sizes are more frequently required to gather data from other countries. This chapter provides those counsel with practical guidance regarding the logistical and operational challenges that arise in a typical matter requiring cross-border discovery.

Chapter 13

Everything in Moderation: Proportionality in Discovery

About this Chapter

In this chapter, we will review each of the six proportionality factors enumerated in FRCP 26(b)(1), some example cases discussing them, and some other considerations. Additionally, we will review guidance from The Sedona Conference on incorporating proportionality considerations into your discovery processes.

13.1 EVERYTHING IN MODERATION

For discovery, the two most significant amendments to the Federal Rules of Civil Procedure ("FRCP") of the last two decades occurred in 2006 and 2015. The 2006 amendments marked the official dawn of the age of eDiscovery, incorporating references to electronically-stored information into the rules and their comments. The 2015 amendments[1] revised, among other things, FRCP 26(b)(1),[2] which defines the scope of discovery. The change brought the existing-but-overlooked concept of proportionality front and center in an attempt to combat the runaway cost and scale of discovery in the digital era.

13.1.1 Before the 2015 Amendments

The pre-amendment version of FRCP 26(b)(1) focused on relevance first and foremost, as well as providing examples of the types of information to which requesting parties are entitled. Later in FRCP 26(b), in a section on limitations applicable to this general discovery scope rule, there was a provision that raised the issue of proportionality. Unfortunately, this limitation was generally overlooked by parties and under-utilized by judges, leading to a lot of disproportional over-discovery ("'. . . the Committee had been told repeatedly that courts were not using these limitations as originally intended'" [3]).The December 2015 amendments attempted to address this issue head on by elevating and emphasizing this proportionality limitation.

13.1.2 After the 2015 Amendments

The 2015 amendments moved the proportionality limitation, from a later subclause, up to the FRCP 26(b)(1) definition of the scope of discovery itself, making it equal in importance to relevance and updating the list of factors to be considered:

> (b) Discovery Scope and Limits.
>
> > (1) Scope in General. Unless otherwise limited by court order, the scope of discovery is as follows: Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense **and proportional to the needs of the case, considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information, the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit**. Information within this scope of discovery need not be admissible in evidence to be discoverable. [emphasis added]

The Committee Notes to the 2015 Amendments[4] explain this is not an entirely new standard but a reemphasized one:

> The present amendment restores the proportionality factors to their original place in defining the scope of discovery. This change reinforces the Rule 26(g) obligation of the parties to consider these factors in making discovery requests, responses, or objections.

[1] Order (U.S. Apr. 29, 2015), available at http://www.supremecourt.gov/orders/courtorders/frcv15(update)_1823.pdf.
[2] Fed. R. Civ. P. 26(b)(1), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[3] Fed. R. Civ. P. 26(b)(1), Committee Notes on Rules—2015 Amendment, available at https://www.law.cornell.edu/rules/frcp/rule_26.
[4] Ibid.

Restoring the proportionality calculation to Rule 26(b)(1) does not change the existing responsibilities of the court and the parties to consider proportionality, and the change does not place on the party seeking discovery the burden of addressing all proportionality considerations.

Nor is the change intended to permit the opposing party to refuse discovery simply by making a boilerplate objection that it is not proportional. The parties and the court have a collective responsibility to consider the proportionality of all discovery and consider it in resolving discovery disputes. [emphasis added]

In general, courts responded to this change by increasing their focus on proportionality and beginning to treat it as a fundamental requirement for obtaining discovery, on par with relevance. Early examples of this shift can be seen in Gilead Sciences[5] ("request is precisely the kind of disproportionate discovery that Rule 26 – old or new – was intended to preclude") and Takata Airbags[6] ("The recently amended Rule 26(b)(1) of the Federal Rules of Civil Procedure 'crystalizes the concept of reasonable limits on discovery through increased reliance on the common-sense concept of proportionality.'").

13.1.3 Proportionality Elsewhere in the Rules

The concept of proportionality also appears elsewhere in the FRCP and in the Federal Rules of Evidence ("FRE"). In the FRCP, it appears in two additional places relevant to discovery:

- First, FRCP 26(b)(2)(B)[7] establishes a specific limitation on the discovery of ESI that is "not reasonably accessible because of undue burden or cost." This is intended to allow for the realities of dealing with legacy systems and storage media. In practice, the analysis of whether there is "undue burden or cost" looks very similar to proportionality analyses under FRCP 26(b)(1).
- Second, FRCP 37(e)[8] addresses what happens when spoliation of ESI occurs "because a party failed to take reasonable steps to preserve it." The Committee Notes from the 2015 amendments make clear that proportionality is a "factor in evaluating the reasonableness of preservation efforts," including consideration of the parties' relative resources.

In the FRE, proportionality is incorporated into the FRE 502(b)[9] analysis associated with determining whether "the holder of the privilege or protection took reasonable steps to prevent disclosure" and "reasonable steps to rectify the error" for the purposes of determining whether the disclosure results in privilege waiver.

[5] *Gilead Sciences, Inc. v. Merck & Co., Inc.*, No. 5:13-cv-04057-BLF (N.D. Cal. Jan. 13, 2016), available at https://docs.justia.com/cases/federal/district-courts/california/candce/5:2013cv04057/269618/211.
[6] *In re Takata Airbag Prods. Liab. Litig.*, MDL No. 2599 (S.D. Fla. Mar. 1, 2016), available at https://ralphlosey.files.wordpress.com/2016/12/in_re_takata_airbag_productionirrelevantfamilymembers.pdf.
[7] Fed. R. Civ. P. 26(b)(2)(B), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[8] Fed. R. Civ. P. 37(e), available at https://www.law.cornell.edu/rules/frcp/rule_37.
[9] Fed. R. Evid. 502(b), available at https://www.law.cornell.edu/rules/fre/rule_502.

13.2 SIX PROPORTIONALITY FACTORS

As noted above, the amended FRCP 26(b)(1) included an updated list of factors to be considered when assessing the proportionality of requested discovery. Those six factors are:

1. The importance of the issues at stake in the action
2. The amount in controversy
3. The parties' relative access to relevant information
4. The parties' resources
5. The importance of the discovery in resolving the issues
6. Whether the burden or expense of the proposed discovery outweighs its likely benefit

13.2.1 The Importance of the Issues at Stake

The first factor encourages consideration of how important the matters under dispute are to the broader legal and social context, to the parties involved, or to potential setting of precedent. For example, a products liability case might have public safety implications, or a class action suit might raise an important question about privacy rights. Even for a purely commercial dispute, the stakes could be economically existential for one or both corporate parties. The more significant the issues are, the more discovery could be deemed proportional.

For an example case discussing this issue, see *First Niagara Risk Mgt., Inc. v. Folino*[10]:

> The issues at stake are of grave importance to First Niagara, who has allegedly uncovered a plan by one of its top executives to start a competing business and employing former First Niagara employees. The first factor therefore weighs in favor of granting First Niagara's motion.

13.2.2 The Amount in Controversy

The second factor considers the monetary stakes of the case. How much does either party stand to win or lose when the case is resolved? The more money at stake, the more money it could be proportional to spend on discovery. For example, it would not make much sense to spend $100,000 on discovery for a legal dispute over only $150,000. Conversely, it would not make much sense to oppose spending $100,000 on discovery for a legal dispute over $1,500,000. As we will discuss further below, courts and parties have tools for tailoring discovery to the right size, including testing, sampling, and phased discovery.

For an example case discussing this issue, see *Oxbow Carbon & Minerals LLC v. Union Pac. R.R.*[11]:

> Here, Oxbow seeks to recover the more than $50,000,000 in illegal fuel surcharges it alleges were the result of the Defendants' collusion. . . . Meanwhile, Oxbow's estimated cost of complying with Defendants' proposed discovery is approximately $140,000 . . . . Given the very substantial amount of damages that Oxbow seeks to recover in this case, its cost of complying with the discovery request to produce information relevant to Defendants' defense of Oxbow's claims does not strike the undersigned as excessive.

[10] *First Niagara Risk Mgt., Inc. v. Folino,* 317 F.R.D. 23 (E.D. Pa. Aug. 11, 2016), available at https://www.ediscoverylaw.com/wp-content/uploads/2016/10/First-Niagra-Risk-Mgmt-Opinion.pdf.
[11] *Oxbow Carbon & Minerals LLC v. Union Pac. R.R.*, Case No. 11-cv-1049 (D.D.C. Sept. 11, 2017), available at https://scholar.google.com/scholar_case?case=9535661083130210393.

13.2.3 The Parties' Relative Information Access

The third factor considers the potential for an informational imbalance to exist in which one party has significantly greater access to relevant materials than the other. Two commercial entities in a dispute are likely to have similar access to relevant materials, but an individual party will often have less access than a corporate one. It may be necessary to impose an unequal discovery burden in order to correct that informational imbalance. The greater the imbalance, the greater the discovery burden it may be proportional to impose.

For an example case discussing this issue, see *Oxbow Carbon & Minerals LLC v. Union Pac. R.R.*[12]:

> In considering this factor, courts look for "information asymmetry"—a circumstance in which one party has very little discoverable information while the other party has vast amounts of discoverable information. . . . Indeed, neither party disputes that Koch is in possession of relevant, unique information, and there appears to be no other way for Defendants to obtain this information than moving to compel Oxbow to produce it.

13.2.4 The Parties' Resources

The fourth factor requires consideration of the financial, technical, and logistical resources available to each party. Discovery can be complex and voluminous, and dramatic differences may exist in the resources available to the parties – particularly when one party is an individual and one is a commercial entity. With the right tools and skills, large volumes of materials can be evaluated efficiently, but without them, it might be prohibitively time-consuming or expensive. The greater the resources available to a party, the greater the discovery burden it may be proportional to place upon them.

For an example case discussing this issue, see *Bourell v. Ronscavage, No. 3:21-CV-01098 (MPS) (D. Conn. June 23, 2023)*[13]:

> *The parties' resources*. With regard to resources . . . Defendants point out that Plaintiff has two law firms representing him with plenty of resources, who are highly competent and capable of handling discovery in a large case. The Court finds that this factor weights in favor of disclosure.

13.2.5 The Importance of the Discovery

The fifth factor requires consideration of how important the particular discovery requested is to resolving the issues. If the requested discovery would be similar to other materials already produced or to other materials that could be more easily or cheaply produced, it might not be important enough to resolving the issues to be proportional. On the other hand, if the requested materials would be unique or potentially significant, it might be important enough to resolving the issues to be proportional. The more the requested discovery would contribute to the efficient resolution of the dispute, the more likely it will be considered proportional.

[12]Ibid.
[13]*Bourell v. Ronscavage*, No. 3:21-CV-01098 (MPS) (D. Conn. June 23, 2023), available at https://app.ediscoveryassistant.com/case_law/50903-bourell-v-ronscavage.
[14]Ibid.
[15]See supra note 10.

For example cases discussing this issue, see *Bourell v. Ronscavage, No. 3:21-CV-01098 (MPS) (D. Conn. June 23, 2023)*[14] and First Niagara Risk Mgt., Inc. v. Folino, 317 F.R.D. 23 (E.D. Pa. Aug. 11, 2016).[15]

13.2.6 The Burden or Expense

This final factor arguably includes all of the above factors within it. It calls for an overall assessment of the "likely benefit" of the requested discovery and "whether the burden or expense" would outweigh it. This factor does not just require consideration of the financial burdens (both direct and indirect) but also of broader considerations. Are there public interests to consider? Could there be a significant effect on a particular market or industry? What about burdens on rights of privacy or confidentiality? The greater the burdens of any kind that the discovery would impose, the greater the likely benefit must be to satisfy the proportionality requirement.

One factor to keep in mind for the future is the ongoing evolution of the discovery tools available to practitioners. The widespread acceptance of technology-assisted review and continuous active learning changed the industry's perception of what's reasonably possible in larger matters. New tools powered by generative AI may precipitate a similar shift in perception[16] over the next few years.

For an example case discussing this issue, see *SinglePoint Direct Solar LLC v. Solar Integrated Roofing Corp.*[17]:

> A voluminous ESI case is always going to be burdensome. This is an unfortunate reality of ESI heavy, high-dollar commercial cases. However, the Court cannot say, given what is at stake, that the burden of document review is so high as to warrant denying Defendants relevant discovery.

13.3 OTHER PROPORTIONALITY ISSUES

In addition to the listed factors, there are other issues it's important for legal practitioner to bear in mind. In particular, practitioners should bear in mind the importance of the ESI protocol, the challenges associated with newer sources, and the importance of specificity in proportionality arguments.

13.3.1 Importance of the ESI Protocol

In addition to considering the six factors enumerated by the rule, courts will typically give great weight to any ESI protocol that the parties negotiated for the case and had formalized via

[16]Cassandre Coyer, "Generative AI and Federal Rules of Civil Procedure: Is It Meant To Be?," LEGALTECH NEWS, https://www.law.com/legaltechnews/2023/10/13/generative-ai-and-federal-rules-of-civil-procedure-is-it-meant-to-be/ (Oct. 13, 2023).
[17]*SinglePoint Direct Solar LLC v. Solar Integrated Roofing Corp.*, No. CV-21-01076-PHX-JAT (D. Ariz. March 21, 2023), available at https://app.ediscoveryassistant.com/case_law/48532-singlepoint-direct-solar-llc-v-solar-integrated-roofing-corp.

court order. Just as contract counterparties are typically held to a bad deal they made, parties opponent in litigation are typically held to a process agreement they made – even if it has produced a marginally disproportionate result.

For an example case discussing this issue, see *McCormick & Co. v. Ryder Integrated Logistics*[18]:

> While the Discovery Order did not march through each of these standards, it clearly took them into account, finding that that the costs of the review were proportional to the needs of the case. . . . **Further—and again—the parties agreed to this review by the plain language of the ESI Protocol.** [emphasis added]

13.3.2 Proportionality of Discovery from Newer Sources

Obtaining discovery from newer source types is an area in which proportionality and burden arguments are often made. Just as recovery of data from a legacy system may be prohibitively difficult or expensive, so to can collection and processing of data from very new systems.

Most recently, these arguments were made about collection and processing of data from collaboration tools like Slack and Teams. As technical solutions for this new challenge became available, courts took that into account in assessing the proportionality of discovery from such tools. For an example case discussing this issue, see *Benebone v. Pet Qwerks, et al*.[19]:

> Based on the evidence presented in the parties' briefing and at the hearing, the Court finds that requiring **review and production of Slack messages by Benebone is generally comparable to requiring search and production of emails** and is not unduly burdensome or disproportional to the needs of this case – if the requests and searches are appropriately limited and focused. [emphasis added]

13.3.3 The Importance of Specificity in Proportionality Arguments

Courts base their proportionality decisions on fact-specific analyses. Consequently, courts are not persuaded by abstract arguments on proportionality. Just claiming a general burden or a hypothetical cost is not sufficient. Instead, arguments need to feature specific details about the materials, about the technical issues, or about the costs and the time required. These specific should be supported by affidavits and other exhibits. Courts have repeatedly declined to accept vague, generalized claims unsupported by specifics.

For an example case discussing this issue, see *Page v. Bragg Comtys., LLC*[20]:

> While Defendants argue that the discovery would be unduly burdensome, cost prohibitive, and harassing, **they have presented nothing to support these assertions. . . . This court has previously rejected unsubstantiated claims** that discovery would pose an undue burden and was not proportional to the needs of the case. [emphasis added]

[18] *McCormick & Co. v. Ryder Integrated Logistics, Inc.*, No. JKB-22-0115 (D. Md. March 08, 2023), available at https://app.ediscoveryassistant.com/case_law/48857-mccormick-co-v-ryder-integrated-logistics-inc.
[19] *Benebone v. Pet Qwerks, et al.*, No. 8:20-cv-00850-AB-AFMx (C.D. Cal. Feb. 18, 2021), available at https://app.ediscoveryassistant.com/case_law/32595-benebone-v-pet-qwerks.
[20] *Page v. Bragg Comtys., LLC,* No. 5:20-CV-336-D (E.D.N.C. Dec. 15, 2022), available at https://app.ediscoveryassistant.com/case_law/46366-page-v-bragg-comtys-llc.

13.4 THE SEDONA CONFERENCE COMMENTARY ON PROPORTIONALITY

13.4.1 The Sedona Conference

After the 2015 amendments to the federal rules reprioritized proportionality, the Sedona Conference decided it was time to revisit the topic to provide updated guidance[21] to practitioners:

> The practical ramifications of including the proportionality factors in the scope of discovery are evolving and many questions remain concerning how practitioners and judges will adjust. Those questions became the main drivers behind the initiative to revisit at this time The Sedona Conference Commentary on Proportionality in Electronic Discovery.

A public comment version was published in November 2016, and the final version was published in May 2017.

13.4.2 The Sedona Conference Principles of Proportionality

The Sedona Conference Commentary enumerates six core principles related to proportionality in eDiscovery:

> Principle 1: The burdens and costs of preserving relevant electronically stored information should be weighed against the potential value and uniqueness of the information when determining the appropriate scope of preservation.
>
> Principle 2: Discovery should focus on the needs of the case and generally be obtained from the most convenient, least burdensome, and least expensive sources.
>
> Principle 3: Undue burden, expense, or delay resulting from a party's action or inaction should be weighed against that party.
>
> Principle 4: The application of proportionality should be based on information rather than speculation.
>
> Principle 5: Nonmonetary factors should be considered in the proportionality analysis.
>
> Principle 6: Technologies to reduce cost and burden should be considered in the proportionality analysis.

For each of these principles, several comments are then provided to explore the meaning and implication of the principle in practice. These comments are annotated with citations to relevant cases and commentaries.

The guidance, in general, is consistent with the case law we reviewed above and the key points from that case law that we identified. The Commentary also provides practical guidance beyond those points, and I want to highlight a few key pieces of that additional guidance on the role of proportionality during preservation, the importance of knowledge at the time, the benefits of early disclosure and dialogue, the application of testing and sampling, and the advantages of phased or iterative discovery.

[21]The Sedona Conference, *Commentary on Proportionality in Electronic Discovery,* 18 SEDONA CONF. J. 141 (2017), available at https://thesedonaconference.org/publication/Commentary_on_Proportionality_in_Electronic_Discovery.

13.4.3 The Role of Proportionality during Preservation

The first Principle in the Commentary addresses how proportionality should be taken into account during preservation, before litigation has commenced and the Federal Rules of Civil Procedure have become applicable. The 2015 Advisory Committee Notes to amended Rule 37(e)[22] suggest that proportionality should be a factor in assessing the reasonableness of pre-litigation preservation efforts. The Sedona Conference Commentary fully endorses this analysis, but it wisely still suggests caution in preserving too narrowly at this early stage of the litigation process:

> It is important to note that in applying principles of proportionality to preservation, **a miscalculation can lead to the permanent loss of relevant information**. In contrast, a miscalculation during production can usually be cured. In particular, at the preservation stage parties should be wary of applying too narrow a definition of what constitutes relevant ESI. [emphasis added]

13.4.4 The Importance of Knowledge at the Time

In discussing the standards to be applied in assessing proportionality and discovery decisions, the Commentary recognizes how parties' understanding of cases evolves over time and emphasizes the importance of assessing decisions after the fact based on the knowledge that was available to the party at the time:

> This analysis should, in turn, depend on the date when the preservation obligation arose and **the knowledge available to that party at the time** when the information was, or could have been, preserved.
>
> . . .
>
> Therefore, a proportional approach to discovery must be measured by **the information available to the parties "as of the time"** requests, responses, or objections are served. A requesting party may lack sufficient information to understand the burden or expense associated with responding to discovery, while a responding party may not fully appreciate the importance of the discovery to the ultimate disposition of the case. [footnotes omitted; emphasis added]

13.4.5 The Benefits of Early Disclosure and Dialogue

Throughout the Commentary, the preservation, cost, and process benefits of early disclosure and dialogue between the parties are repeatedly emphasized, for example:

> Parties often can reduce the risk of loss of relevant information with steps such as the following: (i) **earlier or more complete disclosure** about the substance of their claims and defenses; (ii) communication about the types of information each party considers to be within the duty to preserve . . . .
>
> . . .
>
> Propounding discovery requests at the early stages of the litigation allows parties time to explore compliance with the discovery requests, consider proportionality issues, and bring any disputes before the court for resolution.
>
> . . .

[22] Fed. R. Civ. P. 37(e), Committee Notes on Rules—2015 Amendment, available at https://www.law.cornell.edu/rules/frcp/rule_37.

> **Preliminary steps of this sort** may help the parties agree on cooperative discovery efforts and potentially yield savings by, for example, eliminating the need for some searches or date ranges, identifying custodians, or refining search terms to more effectively target and retrieve relevant information. [emphasis added]

Moreover, Principle 3 states that "[u]ndue burden, expense, or delay resulting from a party's action or inaction should be weighed against that party." As the Comments to that Principle explain:

> Although a party's conduct is not per se a proportionality factor, **failure to engage in early, meaningful discussions designed to develop a discovery plan and avoid potential disputes may properly affect the outcome of any proportionality determination that a court makes**. This is appropriate because a party can be sanctioned for failing "to participate in good faith in developing and submitting a proposed discovery plan as required by Rule 26(f)." [footnote omitted; emphasis added]

13.4.6 The Application of Testing and Sampling

Throughout the Commentary, the many applications and potential benefits of running test searches, reviewing examples, and conducting formalized sampling are all also emphasized:

> In some circumstances, **the courts may order sampling** of the requested information **to determine whether it is sufficiently important** to warrant discovery.
>
> . . .
>
> In addition, **sampling can be used to demonstrate the rate of responsive information, to extrapolate the volume (and therefore costs)** associated with reviewing the potentially responsive ESI. Further, using sampling to demonstrate the rate of responsive information can support an argument that a data source is or is not likely to contain responsive information.
>
> . . .
>
> **Early test searches or early case assessment technology might facilitate agreement on targeting collections or searches** using certain date ranges, platforms or sources, file types, or custodians. In addition, the parties may need to negotiate whether or which search methods might be necessary to further assist in identifying relevant ESI. [footnotes omitted; emphasis added]

13.4.7 The Advantages of Phased or Iterative Discovery

Finally, the Commentary makes the case for approaching discovery in a phased or iterative way to allow for process refinement and revision as the matter progresses and more is learned:

> For these reasons, the court, or the parties on their own initiative, may find it appropriate **to conduct discovery in phases**, starting with discovery of clearly relevant information

available from the most accessible and least expensive sources. . . . Phasing may allow the parties to develop the facts of the case sufficiently to determine how to efficiently and effectively target subsequent discovery. In addition, **phasing discovery may allow the parties to focus first on the information that will be most helpful in assessing litigation risk and facilitating settlement discussions,** or on case-dispositive legal issues that can be decided with minimal factual development. . . . **In short, phased discovery should be viewed as a way to promote the objectives of Rule 1**. [footnotes omitted; emphasis added]

13.5 KEY TAKEAWAYS

There are five key takeaways from this chapter to remember:

1. Since the 2015 amendments to the FRCP, courts and parties have increased their focus on proportionality, beginning to treat it as a fundamental requirement for obtaining discovery, on par with relevance.
2. FRCP 26(b)(1) enumerates a list of six factors to be considered when courts assess the proportionality of requested discovery:
    a. The importance of the issues at stake in the action
    b. The amount in controversy
    c. The parties' relative access to relevant information
    d. The parties' resources
    e. The importance of the discovery in resolving the issues
    f. Whether the burden or expense of the proposed discovery outweighs its likely benefit
3. In addition to those six factors, courts give great weight to any ESI protocol that the parties negotiated and included in the discovery order.
4. When making proportionality arguments to the court, specifics must be provided about the materials, the challenges, and the costs, and those specifics must be supported by affidavits or other evidence.
5. The potential for proportionality disputes can be reduced by following the Sedona Conference's suggestions for early disclosure and dialogue, for liberal use of sampling and testing, and for the negotiation of phased or iterative discovery.

Chapter 14

When the Bough Breaks: Spoliation in eDiscovery

About this Chapter

In this chapter, we will review the analysis established by FRCP 37(e), as well as an assortment of cases applying it, to provide practitioners with the knowledge they need about: what qualifies as reasonable steps to preserve ESI, what qualifies as intent to deprive, and more.

14.1 WHEN THE BOUGH BREAKS

An enormous volume and diversity of electronically-stored information ("ESI") may be relevant to a case, and this diversity and volume can make identification and preservation challenging. It is almost inevitable that some ESI will slip through the cracks. When that happens, Federal Rule of Civil Procedure 37(e)[1] ("FRCP") provides the framework for assessing the loss and its consequences.

This provision was added as part of the 2015 amendments,[2] after a "strikingly, perhaps uniquely, comprehensive and vigorous"[3] public comment period. It sought to bring predictability and consistency to a topic that had been plagued by unpredictability and inconsistent standards across jurisdictions.

14.1.1 FRCP 37(e)

The current subdivision (e) of FRCP 37 reads:

> (e)**Failure to Preserve Electronically Stored Information.** If electronically stored information that should have been preserved in the anticipation or conduct of litigation is lost because a party failed to take reasonable steps to preserve it, and it cannot be restored or replaced through additional discovery, the court:
>
> (1) upon finding prejudice to another party from loss of the information, may order measures no greater than necessary to cure the prejudice; or
>
> (2) only upon finding that the party acted with the intent to deprive another party of the information's use in the litigation may:
>
> (A) presume that the lost information was unfavorable to the party;
>
> (B) instruct the jury that it may or must presume the information was unfavorable to the party; or
>
> (C) dismiss the action or enter a default judgment.

This version of the subdivision made three primary changes from the version that preceded it:

- First, it eliminated old language regarding the "good-faith operation" of a computer system (and the confusion that came with it), and instead focuses on the more straightforward question of whether ESI has been "lost because a party failed to take **reasonable steps to preserve it**" [emphasis added].
- Second, it now requires a showing of **irreplaceability** and **prejudice** before the application of any consequences, and for unintentional losses, it limits those consequences to **curative measures**, thereby reducing the risks associated with minor ESI losses.
- Third, it creates a clear requirement that **intentionality** be found before severe sanctions can be applied (i.e., it adopted the higher of the standards from the pre-existing circuit split).[4]

[1]Fed. R. Civ. P. 37(e), available at https://www.law.cornell.edu/rules/frcp/rule_37.
[2]Order (U.S. Apr. 29, 2015), available at http://www.supremecourt.gov/orders/courtorders/frcv15(update)_1823.pdf.
[3]Report of Advisory Committee on Civil Rules (May 2, 2014), available at http://www.uscourts.gov/rules-policies/archives/committee-reports/advisory-committee-rules-civil-procedure-may-2014.
[4]Ibid.

At a high level, courts' analyses of ESI spoliation issues now generally follow these steps:

1. Had a duty to preserve the ESI arisen?
   - If so, were reasonable steps to preserve it taken?
2. Can the missing ESI be recovered or replaced through additional discovery?
   - If so, what discovery should be directed and who should bear the cost?
3. If it cannot be recovered, does the loss of the ESI prejudice another party?
   - If so, what measures (e.g., procedural, evidentiary) would cure the prejudice?
4. Is there evidence that spoliating party acted with intent to deprive another party of it?
   - If so, are adverse inferences, dismissal or other severe sanctions warranted?

The points in this analysis where disputes most frequently arise are whether the steps taken to preserve were reasonable and whether there is evidence of intent to deprive.

14.2 REASONABLE STEPS TO PRESERVE ESI

FRCP 37(e) limits the application of sanctions to situations where ESI that should have been preserved was lost "because a party failed to take reasonable steps to preserve it." The rule itself does not elaborate on what qualifies as reasonable steps, but the Advisory Committee Notes to the 2015 Amendments do provide some guidance.

14.2.1 Five Factors to Consider

First and most importantly, the notes emphasize several times that, "[t]his rule recognizes that 'reasonable steps' to preserve suffice; it does not call for perfection." In addition to reemphasizing this general principle of discovery, the Notes also provide a list of five specific factors that courts should consider when performing a post-hoc assessment of whether the steps taken in a given case were "reasonable":

1. The first factor is essentially the prior version of Rule 37(e), which had attempted to provide a narrow safe harbor for ESI loss:
   a. "As under the current rule, the routine, good-faith operation of an electronic information system would be a relevant factor for the court to consider . . . although the prospect of litigation may call for reasonable steps to preserve information by intervening in that routine operation."
2. The second factor is akin to a force majeure clause in a contract that allows for uncontrollable outside events, although reasonable preventative measures may still be expected of parties (e.g., maintaining backups):
   a. ". . . information the party has preserved may be destroyed by events outside the party's control — the computer room may be flooded, a 'cloud' service may fail, a malign software attack may disrupt a storage system, and so on. Courts

may, however, need to assess the extent to which a party knew of and protected against such risks."

3. The third factor courts are directed to consider is the relative sophistication of the parties, particularly with regard to the likely difference in sophistication between large organizations and individuals.
4. The fourth factor courts are directed to consider is the relative resources available to the parties, including financial and human resources. The Notes explicitly state that less-expensive but substantially-as-effective alternatives can be reasonable:

    a. "The court should be sensitive to party resources; aggressive preservation efforts can be extremely costly, and parties (including governmental parties) may have limited staff and resources to devote to those efforts. A party may act reasonably by choosing a less costly form of information preservation, if it is substantially as effective as more costly forms."

5. The final factor that courts are directed to consider is proportionality itself, which is a foundational requirement for all discovery (and which implicates another multi-factor analysis similar to this one).

14.2.2 Decisions Discussing Reasonable Steps

A variety of courts have had the opportunity to issue orders on motions for spoliation sanctions and to consider whether a party had taken the required reasonable steps. Here is a sampling of those cases from the past five years:

- *Paisley Park Enters., Inc. v. George Ian Boxill, Rogue Music Alliance, LLC*, 330 F.R.D. 226 (D. Minn. 2019)[5] – failure to "suspend the auto-erase function on their phones" or to "put in place a litigation hold to ensure that they preserved text messages" found not to have been reasonable steps to preserve
- *Nuvasive, Inc. v. Kormanis*, No. 1:18CV282 (M.D. N.C. Mar. 13, 2019)[6] – failure to "investigate[] . . . what text messages his iPhone held, and [] whether any setting on his iPhone might cause the deletion of existing or future text messages" and failure to "obtain[] appropriate advice about saving back-up copies of his text messages" found not to have been reasonable steps to preserve
- *DriveTime Car Sales Company, LLC v. Pettigrew*, No. 2:17-cv-371 (S.D. Ohio Apr. 18, 2019)[7] – failure to preserve text messages prior to replacing phone with a new one found not to have been reasonable steps to preserve
- *Cruz v. G-Star Inc.*, 17-CV-7685 (PGG) (OTW) (S.D.N.Y. Jun. 19, 2019)[8] – failure to issue a timely legal hold or monitor hold compliance, leading to deletion of emails and SAP data, found not to have been reasonable steps to preserve

[5] *Paisley Park Enters., Inc. v. George Ian Boxill, Rogue Music Alliance, LLC*, 330 F.R.D. 226 (D. Minn. 2019), available at https://casetext.com/case/paisley-park-enters-inc-v-george-ian-boxill-rogue-music-alliance-llc-1.
[6] *Nuvasive, Inc. v. Kormanis*, No. 1:18CV282 (M.D. N.C. Mar. 13, 2019), available at https://casetext.com/case/nuvasive-inc-v-kormanis.
[7] *DriveTime Car Sales Company, LLC v. Pettigrew*, No. 2:17-cv-371 (S.D. Ohio Apr. 18, 2019), available at https://casetext.com/case/drivetime-car-sales-co-v-pettigrew-1.
[8] Cruz v. G-Star Inc., 17-CV-7685 (PGG) (OTW) (S.D.N.Y. Jun. 19, 2019), available at https://casetext.com/case/cruz-v-g-star-inc.

- *In re Google Play Store Antitrust Litig.*, No. 21-md-02981-JD (N.D. Cal. March 28, 2023)[9] – failure to suspend automated janitorial functions, giving "each employee carte blanche to make his or her own call about what might be relevant," and "intentionally deciding not to check up on employee decisions" (essentially "a 'don't ask, don't tell' policy for Chat preservation") found not to have been reasonable steps to preserve

This collection of cases suggests that what qualifies as reasonable steps includes steps like: issuing timely legal holds, monitoring compliance with those holds, suspending automated janitorial functions, and preserving through preemptive collection when needed.

It should also be noted here that the range of sources to which these expectations apply continues to expand. In recent years, parties have been sanctioned for spoliation of Slack messages, Google Vault documents, Basecamp project files, ephemeral Signal messages, and more (e.g., *Red Wolf Energy Trading, LLC v. BIA Cap. Mgmt., LLC*[10]; *Drips Holdings, LLC v. Teledrip LLC*[11]; *Ace Am. Ins. Co. v. First Call Envtl.*[12]; and *Fed. Trade Comm'n v. Noland*[13]). To avoid inadvertent spoliation due to simple lack of awareness, it's important to stay informed about newer tools and communication options that may become sources of relevant ESI if your clients or custodians have chosen to use them.

14.3 INTENT TO DEPRIVE

Before the amendments to FRCP 37(e), a circuit split had arisen regarding the question of whether adverse inference instructions or dismissal could be based merely on some level of negligence or if intentional misconduct was required. As noted above, one of the things the amendments were intended to do was to resolve that split, because the jurisdictional variations created uncertainty for litigants and, allegedly, increased preservation costs. The Advisory Committee on Civil Rules explained in its May 2014 Report[14]:

> Some circuits, like the Second, hold that adverse inference jury instructions (viewed by most as a serious sanction) can be imposed for the negligent or grossly negligent loss of ESI. Other circuits, like the Tenth, require a showing of bad faith before adverse inference instructions can be given. The public comments credibly demonstrate that persons and entities over-preserve ESI out of fear that some might be lost, their actions with hindsight might be viewed as negligent, and they might be sued in a circuit that permits adverse inference instructions or other serious sanctions on the basis of negligence.

[9] *In re Google Play Store Antitrust Litig.*, No. 21-md-02981-JD (N.D. Cal. March 28, 2023), available at https://storage.courtlistener.com/recap/gov.uscourts.cand.373179/gov.uscourts.cand.373179.469.0.pdf.
[10] *Red Wolf Energy Trading, LLC v. BIA Cap. Mgmt., LLC*, 2022 WL 4112081 (D. Mass. Sept. 8, 2022), available at https://app.ediscoveryassistant.com/case_law/44507-red-wolf-energy-trading-llc-v-bia-capital-mgmt-llc.
[11] *Drips Holdings, LLC v. Teledrip LLC*, Sept. 8, 2022 WL 4545233 (N.D. Ohio Sept. 29, 2022), available at https://app.ediscoveryassistant.com/case_law/44997-drips-holdings-llc-v-teledrip-llc.
[12] *Ace Am. Ins. Co. v. First Call Envtl.*, LLC, 2023 WL 137456 (E.D. Pa. Jan. 9, 2023), available at https://app.ediscoveryassistant.com/case_law/46858-ace-am-ins-co-v-first-call-envtl-llc.
[13] *Fed. Trade Comm'n v. Noland*, No. CV-20-00047-PHX-DWL (D. Ariz. Aug. 30, 2021), available at https://app.ediscoveryassistant.com/case_law/36010-fed-trade-comm-n-v-noland.
[14] Report of Advisory Committee on Civil Rules, *supra* note 3.

The amended version of FRCP 37(e)(2) resolved this split in favor of the higher standard by requiring a showing that "the party acted with the intent to deprive another party of the information's use in the litigation" for the application of adverse inference instruction, dismissal, or default judgment sanctions.

Although this rule change increased predictability and reduced the frequency of severe spoliation sanctions, it created new ambiguity around what level of evidence is sufficient to support a finding of intent to deprive. Can intent be inferred or is direct evidence needed?

14.3.1 Decisions Discussing Intent to Deprive

A variety of courts have had the opportunity to issue orders on motions for spoliation sanctions considering whether a party had "acted with the intent to deprive another party of the information's use in the litigation." Here is a sampling of those cases from the past five years:

- *DriveTime Car Sales Company, LLC v. Pettigrew*, No. 2:17-cv-371 (S.D. Ohio Apr. 18, 2019)[15] – in this case, a party failed to take reasonable steps to preserve relevant ESI, but no evidence of intention to deprive beyond the failure itself was shown, leading to no finding of intent to deprive under FRCP 37(e)(2)
- *GN Netcom, Inc. v. Plantronics, Inc.*, 930 F.3d 76 (3rd Cir. 2019)[16] – in this case, "the District Court reasonably concluded that Plantronics acted in bad faith" based on the facts that a Senior VP "deliberately deleted an unknown number of emails in response to 'pending litigation' and urged others to do the same," that "executives, including its CEO, were not truthful during depositions," and that "the company was not willing to spend a nominal fee for its expert, Stroz, to fully assess the spoliation and create a final report," each of which "was an intentional step to interfere with GN's prosecution of its claims against Plantronics"
- *GMS Indus. Supply, Inc. v. G&S Supply, LLC*, No. 2:19-cv-324 (RCY) (E.D. Va. Mar. 22, 2022)[17] – in this case, the court found intent to deprive based on the defendant's decision, after receiving hold notices, to download an application called File Shredder and use it to permanently delete all user created files on his computer
- *Jennings v. Frostburg State Univ.*, No. ELH-21-656 (D. Md. June 27, 2023)[18] – in this case, the court declined to infer intent to deprive from defendant's failure to implement a litigation hold in a timely manner or from the erasure of two relevant custodians' phones when their employment ended
- *Skanska USA Civil Se. Inc. v. Bagelheads, Inc.*, 75 F.4th 1290 (11th Cir. 2023)[19] – in this case, the Second Circuit affirmed a finding of intent to deprive that was based on a party's systemic preservation failures without more direct evidence of bad faith:

> The court found a "lack of any cogent explanation" for Skanska's complete failure to make any effort to preserve the destroyed cell phones. It focused in particular on how the company "took no action" to educate its custodians and administrators about the litigation hold and "made no effort" to collect its custodians' cell phone data until at least seven months after the litigation hold was in place. . . . In the district court's view, bad faith was the only thing that explained the company's actions. [internal citations omitted]

[15] *DriveTime Car Sales Company, supra* note 7.
[16] *GN Netcom, Inc. v. Plantronics, Inc.*, 930 F.3d 76 (3rd Cir. 2019), available at https://www2.ca3.uscourts.gov/opinarch/181287p.pdf.
[17] *GMS Indus. Supply, Inc. v. G&S Supply, LLC*, No. 2:19-cv-324 (RCY) (E.D. Va. Mar. 22, 2022), available at https://app.ediscoveryassistant.com/case_law/40609-gms-indus-supply-inc-v-g-s-supply-llc.
[18] *Jennings v. Frostburg State Univ.*, No. ELH-21-656 (D. Md. June 27, 2023), available at https://app.ediscoveryassistant.com/case_law/50818-jennings-v-frostburg-state-univ.
[19] *Skanska USA Civil Se. Inc. v. Bagelheads, Inc.*, 75 F.4th 1290 (11th Cir. 2023), available at https://casetext.com/case/skanska-us-civil-se-v-bagelheads-inc.

> The court noted however, that were its review de novo rather than for clear error, it would have been a "close question" whether to find bad faith or merely gross negligence.

These decisions considering intent to deprive show a bit more variation than the decisions we reviewed regarding reasonable steps. As with many aspects of discovery, courts' decisions in these cases are very fact-specific. In general, courts seem reluctant to infer intent solely from a lack of reasonable steps or other circumstantial evidence, but some are willing to draw that inference – particularly when the failures have been egregious or the explanations implausible.

14.4 OTHER FACTORS TO REMEMBER

In addition to reasonable steps and intent to deprive, there are other factors related to spoliation sanctions worth remembering. First, there must have been irretrievable loss of some ESI for sanctions to be applied under the rule. Second, there must also have been prejudice to another party from that loss. Third, sanctions need not be proportional to the value of the case. Finally, courts are not always limited to the terms of the rule.

14.4.1 There Must Have Been Irretrievable Loss

The language of FRCP 37(e) specifies that, in order to apply sanctions under the rule, ESI that should have been preserved has been "lost . . . and [] cannot be restored or replaced through additional discovery." So, even in cases where reasonable steps to preserve weren't taken, sanctions may not apply if no ESI was lost or if any lost ESI can be recovered or replaced (e.g., *Globus Med., Inc. v. Jamison*[20]). The availability of alternate forms of the lost evidence (e.g., screen captures, testimony) or alternate sources of the lost evidence (e.g., third parties or service providers) may be sufficient to preclude a finding of loss (e.g., *Envy Hawaii LLC v. Volvo Car USA LLC*[21]), but alternate forms (including screen captures) are not always sufficient (e.g., *Edwards v. 4JLJ, LLC*[22]).

It should also be noted here that, technically, you can have spoliation without irretrievable loss in those rare cases where someone has altered or fabricated evidence. Those too are sanctionable forms of spoliation, and such intentional misconduct receives the harshest penalties (e.g., *Gunter v. Alutiiq Advanced Sec. Sols., LLC*[23]; *Rossbach v. Montefiore Med. Ctr.*[24]).

14.4.2 There Must Have Been Prejudice

The language of FRCP 37(e) also specifies that, in order to apply even curative measures under the rule, the court must find "prejudice to another party from loss of the information." So, even in cases where reasonable steps to preserve weren't taken and ESI was irretrievably lost, sanctions may not apply if no prejudice from the loss can be shown (e.g., Hernandez v. Tulare County Correction Center[25]; Sinclair v. Cambria Cnty.[26]).

[20] *Globus Med., Inc. v. Jamison,* 2023 WL 2127410 (E.D. Va. Feb. 10, 2023), available at https://app.ediscoveryassistant.com/case_law/47832-globus-med-inc-v-jamison.
[21] *Envy Hawaii LLC v. Volvo Car USA LLC,* No. 17-00040 HG-RT (D. Haw. Mar. 20, 2019), available at https://scholar.google.com/scholar_case?case=13587045757304291469.
[22] *Edwards v. 4JLJ, LLC,* No. 2:15-CV-299 (S.D. Tex. Jan. 11, 2019), available at https://casetext.com/case/edwards-v-4jlj-llc-2.
[23] *Gunter v. Alutiiq Advanced Sec. Sols., LLC,* No. 1:20-CV-03410-JRR (D. Md. March 2, 2023), available at https://app.ediscoveryassistant.com/case_law/48019-gunter-v-alutiiq-advanced-sec-sols-llc.
[24] *Rossbach v. Montefiore Med. Ctr.*, No. 21-2084 (2d Cir. August 28, 2023), available at https://app.ediscoveryassistant.com/case_law/52134-rossbach-v-montefiore-med-ctr.

14.4.3 Sanctions Can Exceed Case Value

It is important to remember that discovery sanctions need to be pegged to the prejudice they are curing or the conduct they are deterring rather than to the value of the overall case:

- *Klipsch Grp., Inc. v. ePRO E-Commerce Ltd.*, 880 F.3d 620 (2d Cir. 2018)[27] – in this case, the Second Circuit approved discovery sanctions – including a $2.7 million award of fees and costs in a case with a value of around $20,000 – over the Defendant's objection that such sanctions were "impermissibly punitive, primarily because they are disproportionate to the likely value of the case," because as the court explained:

> . . . discovery sanctions should be commensurate with the costs unnecessarily created by the sanctionable behavior. A monetary sanction in the amount of the cost of discovery efforts that appeared to be reasonable to undertake ex ante does not become impermissibly punitive simply because those efforts did not ultimately uncover more significant spoliation and fraud, or increase the likely damages in the underlying case.

14.4.4 Courts Are Not Limited to FRCP 37(e)

One of the primary goals of the December 2015 Amendments to the Federal Rules of Civil Procedure was to increase predictability and consistency for litigants by eliminating jurisdictional variations in ESI spoliation standards, their application, and the associated penalties. To accomplish the desired standardization, amended FRCP 37(e) would need to become the only source of authority for the application of ESI spoliation sanctions, to prevent the rule from being bypassed and predictability from being destroyed.

As articulated in the Advisory Committee Notes to the 2015 Amendments,[28] the Rules Advisory Committee intended for the new version of FRCP 37(e) to preclude the use of inherent authority to assess ESI spoliation sanctions:

> New Rule 37(e) . . . authorizes and specifies measures a court may employ if information that should have been preserved is lost, and specifies the findings necessary to justify these measures. It therefore forecloses reliance on inherent authority or state law to determine when certain measures should be used.

Advisory notes are only advisory, however, and in the years since, courts have generally followed the rule but continued to assert the inherent power to sanction as needed to manage their proceedings and ensure just resolutions. For example, in *CAT3 LLC v. Black Lineage*,[29] the court imposed sanctions under Rule 37(e) but asserted explicitly that inherent authority would still have been an option for imposing the sanctions, if the result provided by the rule had been inadequate:

> Where exercise of inherent power is necessary to remedy abuse of the judicial process, it matters not whether there might be another source of authority that could address the same issue. In Chambers, the Supreme Court rejected the argument by the party opposing the sanctions motion that provisions of the Federal Rules of Civil Procedure foreclosed resort to inherent power. It stated that "the inherent power of a court can

[25] *Hernandez v. Tulare County Correction Center,* No. 16-CV-00413 (E.D. Cal. Feb. 8, 2018), available at https://casetext.com/case/hernandez-v-tulare-cnty-corr-ctr-2.
[26] *Sinclair v. Cambria Cnty.*, No. 3:17-cv-149 (W.D. Pa. Sept. 28, 2018), available at https://scholar.google.com/scholar_case?case=8414468043782276336.
[27] *Klipsch Grp., Inc. v. ePRO E-Commerce Ltd.*, 880 F.3d 620 (2d Cir. 2018), available at https://casetext.com/case/klipsch-grp-inc-v-epro-e-commerce-ltd-1.
[28] Fed. R. Civ. P. 37(e), Committee Notes on Rules—2015 Amendment, available at https://www.law.cornell.edu/rules/frcp/rule_37.
[29] *CAT3 LLC v. Black Lineage,* 164 F. Supp. 3d 488 (S.D.N.Y. Jan. 12, 2016), available at https://casetext.com/case/cat3-llc-v-black-lineage-inc-2.

> be invoked even if procedural rules exist which sanction the same conduct." [internal citations omitted]

Moreover, this inherent authority also gives courts great discretion in fashioning appropriate remedies and sanctions. For example, in combination with more common sanctions like fee awards and adverse inference instructions, a court might also shift a burden of proof (e.g., *Edwards v. 4JLJ, LLC*[30]) or preclude the calling of specific witnesses (e.g., *Wilmoth v. Deputy Austin Murphy*[31]).

14.5 KEY TAKEAWAYS

There are five key takeaways from this chapter to remember:

1. Under FRCP 37(e), reasonable steps to preserve ESI includes steps like: issuing timely legal holds, monitoring compliance with those holds, suspending automated janitorial functions, and preserving through preemptive collection when needed.
2. If the prejudicial loss was accidental, only curative measures can be imposed, but if there was intent to deprive another party, severe sanctions can be imposed.
3. Some courts decline to infer intent solely from a lack of reasonable steps or other circumstantial evidence, but some are willing to draw that inference – particularly when the failures have been egregious or the explanations implausible.
4. Even in cases where reasonable steps to preserve weren't taken, sanctions may not apply if the ESI can be recovered elsewhere or if no prejudice from the loss can be shown based on what the contents of the missing ESI would have been.
5. When necessary to manage their affairs or to craft an appropriate remedy, courts may go beyond the letter of FRCP 37(e) and rely on their inherent authority to apply sanctions at a time or of a type not dictated by the rule.

[30] *Edwards, supra* note 22.
[31] *Wilmoth v. Deputy Austin Murphy,* No. 5:16-CV-5244 (W.D. Ark. Aug. 7, 2019), available at https://casetext.com/case/wilmoth-v-murphy-1.

Chapter 15

An Ounce of Prevention: Fundamentals of Data Protection

About this Chapter

In this chapter, we will discuss why ESI must be protected and how you can protect it, including fundamentals of security compliance frameworks, role-based access control, cloud storage vs. on-premises storage, and data encryption.

15.1 INTRODUCTION

It is axiomatic that the volumes of electronically-stored information (ESI) generated by organizations are vast and ever-increasing. Correspondingly, the amount of ESI that must be preserved, collected, processed, and reviewed for internal investigations, active litigation, and regulatory compliance never stops growing. Today, it is a practical and ethical requirement for practitioners in these areas to take the necessary steps to protect the ESI they are managing for those purposes, which means keeping up with evolving security and compliance best practices – as well as adapting to the rapidly changing tactics to threat actors.

15.2 WHY ESI MUST BE PROTECTED

So, why is data protection something lawyers and other legal practitioners need to know? Isn't that IT's job? The short answer is that data protection is everyone's job, and the long answer is that data protection in the legal industry is critical both because of lawyers' ethical duties and because of the potential consequences of a breach into such sensitive data.

15.2.1 Lawyers' Ethical Duties

In August 2012, the American Bar Association (ABA) implemented changes[1] to its Model Rules of Professional Conduct, including a change to make the need for technology competence explicit. In the eleven years since the change to the Model was implemented, forty states have adopted some form of this technology competence requirement for lawyers.[2] Although this change was spurred in large part by the rapid rise of eDiscovery, it is not limited to just that area.[3] It encompasses technology competence in several contexts, including "safeguarding client information" and "the technology that lawyers use to run their practices."

In addition to the duty of technology competence, lawyers have an ethical duty to protect client confidentiality. For example, ABA Model Rule of Professional Conduct 1.6(c)[3] says that "[a] lawyer shall make reasonable efforts to prevent the inadvertent or unauthorized disclosure of, or unauthorized access to, information relating to the representation of a client." The ABA has elaborated on this duty in the contexts of cybersecurity and remote work:

- ABA Formal Opinion 477R[5] (2017) discusses lawyers' obligation to understand and use electronic security measures to safeguard client communications and information. The opinion discusses a range of best practices to employ, including: using strong, unique passwords; enabling multifactor authentication; securing Wi-Fi networks; updating software regularly; and enabling antivirus software and firewalls.
- ABA Formal Opinion 498[6] (2021) discusses lawyers' obligations to safeguard information when working remotely. The opinion discusses a range of best practices that overlaps with the recommendations from Formal Opinion 477R and

[1]Debra Cassens Weiss, *Lawyers Have Duty to Stay Current on Technology's Risks and Benefits, New Model Ethics Comment Says*, ABA JOURNAL, http://www.abajournal.com/news/article/lawyers_have_duty_to_stay_current_on_technologys_risks_and_benefits/ (Aug. 6, 2012).
[2]Robert Ambrogi, *Tech Competence*, LAWSITES, https://www.lawsitesblog.com/tech-competence (last visited Dec. 20, 2023).
[3]Steven M. Puiszis, Perspective: *Technology Brings a New Definition of Competency*, BLOOMBERG LAW, https://news.bloomberglaw.com/business-and-practice/perspective-technology-brings-a-new-definition-of-competency (Apr. 12, 2016).
[4]ABA Model Rules of Prof'l Conduct R. 1.6 (2023), available at https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_6_confidentiality_of_information/.
[5]"ABA Formal Opinion 477R: Securing communication of protected client information," American Bar Association (June 2017), available at https://www.americanbar.org/news/abanews/publications/youraba/2017/june-2017/aba-formal-opinion-477r--securing-communication-of-protected-cli/
[6]"ABA issues guidance on model rules, ethical tech duties to consider when working remotely," American Bar Association (Mar. 10, 2023), available at https://www.americanbar.org/news/abanews/aba-news-archives/2021/03/aba-issues-guidance-on-model-rules--ethical-tech-duties-to-consi/.

adds some additional recommendations related to the use of encryption, virtual private networks (VPNs), data backups, breach policies, and more.

15.2.2 Potential Consequences of a Breach

Beyond lawyers' ethical duties, there are also a range of potential consequences of data breach, loss, or exposure that range from embarrassing to costly to criminal. Any pool of ESI collected for discovery or investigation may contain not only privileged materials but also a wide range of other sensitive materials such as personally identifiable information (PII), personal health information (PHI), customer information, proprietary information, trade secrets, and even classified information.

Your organization may be obligated to protect such information by court rules, by contract, by federal and state law, and by international law. For example, in the U.S., disclosure of personally-identifiable medical information generally needs to be prevented to comply with the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPPA).[7] When dealing with ESI in the EU, disclosure of personally-identifiable information may need to be prevented to comply with the General Data Protection Regulation (GDPR).[8]

Failing to prevent a data breach or other inadvertent disclosure can obviously result in privilege waiver, but it can also create an obligation to report the breach to regulators and individuals, result in a loss of trade secret status, or cause significant reputational damage – and, of course, investigating and remediating the issue is also likely to be expensive and time-consuming.

15.3 HOW TO PROTECT ESI

In order to reduce the potential for those negative outcomes, legal and compliance practitioners responsible for managing ESI need to understand the fundamentals of security compliance frameworks, role-based access control, cloud storage vs. on-premises storage, data encryption, and adaptation to evolving adversarial tactics.

15.3.1 Security Compliance Frameworks: ISO, SOC2, and NIST

Security compliance frameworks play a crucial role in ensuring the protection of data and the security of eDiscovery processes, and selecting service providers that adhere to them will help ensure the protection of the ESI you are managing. There are three notable frameworks that are often applied in the legal industry: ISO 27001, SOC2, and NIST SP 800-53.

1. **ISO 27001/27002 (International Organization for Standardization)**

   ISO 27001/27002 are internationally recognized standards for information security management systems (ISMS). In the eDiscovery industry, compliance with ISO 27001/27002 demonstrates a systematic approach to managing sensitive information, reducing risks, and ensuring data security. While ISO 27001 defines the standards an organizations are required to meet, 27002 offers a comprehensive

[7]U.S. Dep't of Health & Hum. Servs., *Summary of the HIPAA Privacy Rule*, HHS.gov, https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html (July 26, 2013).
[8]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 59, 1 (May 4, 2016), available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504&qid=1532348683434.

set of controls that include policies, procedures, technical measures, and employee training. eDiscovery providers that adhere to ISO 27001/27002 standards are indicating a commitment to safeguarding client data that includes support from the top of the organization down.

2. **SOC2 (Service Organization Control)**

   SOC2 is a set of auditing standards developed by the American Institute of CPAs (AICPA) for service organizations. It focuses on security, availability, processing integrity, confidentiality, and privacy. In the eDiscovery context, SOC2 compliance is particularly relevant as it assesses the effectiveness of an organization's controls over data security and privacy. When reviewing a provider's overall security program, it is important to identify whether they have obtained a SOC2 Type 1 or SOC2 Type 2 certification. Type 1 certification means that the security controls the provider has in place were reviewed at a single point in time, while the more common Type 2 certification shows that the effectiveness of those security controls were assessed over a period of time. eDiscovery providers that obtain SOC2 compliance demonstrate their commitment to meeting high standards of security and privacy.

3. **NIST SP 800-53 (National Institute of Standards and Technology)**

   NIST SP 800-53 is a comprehensive framework that provides a catalog of security controls for U.S. Federal Government information systems and organizations. While originally developed for the U.S. government, it has been widely adopted in the private sector, particularly among those organizations that provide software and/or services to the Federal Government. The Federal Risk and Authorization Management Program, more commonly known as FedRAMP, draws its security controls directly from this publication, and service providers who wish to handle sensitive data should be knowledgeable of them.

15.3.2 Role-Based Access Control

Role-Based Access Control (RBAC) is a fundamental principle of data protection in the eDiscovery industry. It is a security model that restricts system access to authorized users based on their job-specific roles within the organization. RBAC ensures that individuals only have access to the data and resources necessary for them to perform their job functions. This not only enhances security but more importantly helps maintain the "principle of least privilege," which reduces the risk of data breaches by ensuring users only have access to the ESI and resources they need to complete their required tasks.

In eDiscovery, RBAC is essential for controlling access to sensitive legal and case-related information. Different personnel, such as lawyers, paralegals, IT administrators, and external counsel, should have distinct roles with corresponding access permissions. For example, a lawyer might have full access to case files and legal documents, while an IT administrator may only have access to system configuration settings. Implementing RBAC helps organizations adhere to data protection and privacy regulations, like the GDPR in Europe and HIPAA in the United States.

15.3.3 Data Storage: Cloud vs. On-Premises

Most of the data implicated in eDiscovery or compliance activities is highly sensitive in nature. The choice of whether to store that data on-premises (often in a data center) or to allow that data to be stored in the cloud is a critical decision with significant implications for data protection and security, both when considering your own organization's storage and when considering the storage used by any legal services providers with whom you work.

- **Cloud Storage**

  Public cloud providers like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud have become increasingly popular options in which to store massive amounts of data. Data storage in the cloud offers numerous advantages, such as massive scalability for large-scale datasets and an ease of accessibility, as the data can be accessed from any device that is connected to the Internet. However, data protection in the cloud must be approached with care. Legal practitioners should evaluate the security controls in place with cloud-based data storage providers, in particular with respect to access management controls to ensure that the data is only available to those who should have access. There have been multiple instances of cloud data repositories being accessible to the world due to misconfigurations in their access management settings. Additional considerations include encryption of data in transit to get to the cloud storage repository and the potential access by storage administrators from the cloud storage provider.

- **On-Premises Storage**

  On-premises data storage provides organizations with a greater amount of control over their data, including knowing exactly where their data is physically located and where it is is backed up. This approach is ideal for organizations with specific regulatory requirements or heightened security requirements, as it allows for full control of the environment in which sensitive data will be used. Additionally, where certain types of projects require special or additional security controls to be applied, having full control of the environment provides the ability for a more granular approach to security.

Hybrid approaches, combining the benefits of cloud and on-premises storage, have gained popularity in the eDiscovery industry. This allows organizations to maintain sensitive data on-premises while using the cloud for backup, collaboration, and data processing. Ultimately, the choice between cloud and on-premises storage should align with an organization's specific security, compliance, and operational requirements.

15.3.4 Data Encryption

Data encryption is a cornerstone of data protection, not only in the eDiscovery industry, but globally. It ensures that sensitive information remains confidential and secure, even in the event of unauthorized access. There are two contexts in which the encryption of your ESI must be considered: encryption at rest and encryption in transit.

- **Encryption at Rest**
  - Encryption at rest protects data stored on physical and digital media, such as hard drives, servers, and databases. In the context of eDiscovery, this means that case files, legal documents, and client data should be stored in an encrypted format. The encryption keys should be securely managed and

stored separately from the data to prevent unauthorized decryption.

- Common encryption techniques for data at rest include full disk encryption (FDE) and file-level encryption. Many eDiscovery providers rely on industry-standard encryption algorithms like AES (Advanced Encryption Standard) to protect their stored data. Additionally, organizations may opt for hardware security modules (HSMs) included as part of large storage arrays to enhance the security of encryption key management.

- **Encryption in Transit**
  - Encryption in transit ensures that data remains secure while it is being transmitted over networks, including the internet. Legal practitioners and eDiscovery providers often exchange sensitive information with clients, partners, and regulatory bodies. Therefore, it is essential to use secure communication protocols such as SSL/TLS to encrypt data during transmission. Secure email and file transfer solutions are also commonly employed in eDiscovery to protect the confidentiality of messages and attachments.
  - The integration of end-to-end encryption (E2EE) can add an extra layer of protection, ensuring that only the intended recipients can decrypt and access the data. E2EE is particularly crucial when transmitting highly sensitive legal documents and case-related information.

15.3.5 Adapting to Ransomware Threats

The legal industry is certainly not immune to the evolving landscape of ransomware threats. In recent years, cybercriminals have increasingly targeted organizations across all sectors, including law firms and eDiscovery providers, with ransomware attacks. These attacks can have severe consequences, including data breaches, data loss, legal liabilities, and significant financial losses.

To defend against ransomware threats, organizations must establish robust security programs that continually evolve to match changing adversarial tactics. Here are some key strategies for mitigating ransomware risks:

1. **Regular Backup and Recovery Procedures**

   Regularly back up critical data and test the restoration process to ensure it is effective. Backup data should be stored in a secure and isolated location, away from the primary network, to prevent ransomware attacks from compromising the backups.

2. **Employee Training and Awareness**

   Invest in cybersecurity training and awareness programs for employees to help them recognize phishing attempts and other social engineering tactics commonly used by ransomware attackers. A well-informed workforce is a critical defense against these threats.

3. **Network Segmentation**

   Implement network segmentation to isolate critical systems and data from less sensitive areas of the network. If a ransomware attack occurs, this can help contain the impact and prevent lateral movement within the network.

4. **Patch Management**

   Maintain up-to-date software and systems with the latest security patches. Many

ransomware attacks exploit known vulnerabilities in outdated software.

5. **Advanced Endpoint Detection and Response (EDR) Solutions**

   Deploy EDR solutions that can detect and respond to unusual or suspicious activity on endpoints. These solutions can help identify and halt ransomware attacks in progress.

6. **Threat Intelligence Sharing**

   Participate in threat intelligence sharing communities, such as the Legal Services Information Sharing and Analysis Organization (LS-ISAO), and share information about emerging threats. Collaborating with industry peers and security organizations can provide valuable insights into the latest tactics used by ransomware actors.

7. **Incident Response Plans**

   Develop and regularly update incident response plans to facilitate a coordinated and effective response in the event of a ransomware attack. This includes procedures for containment, eradication, and recovery.

15.4 CONCLUSION

For legal practitioners, a commitment to data protection is a matter of ethical duty, of legal and regulatory compliance, and of fundamental responsibility to clients and stakeholders. As the digital landscape continues to evolve, staying ahead of security challenges and mitigating risks must remain a top priority for all legal and compliance professionals. Doing so requires consideration of security compliance frameworks, role-based access control, cloud storage vs. on-premises storage, and data encryption.

15.5 KEY TAKEAWAYS

There are five key takeaways from this chapter to remember:

1. Data protection is the job of everyone within an organization, and data protection in the legal industry is critical both because of lawyers' ethical duties and because of the potential consequences of a breach into sensitive data.
2. Security compliance frameworks like ISO 27001, SOC2, and NIST SP 800-53 provide a structured approach to safeguarding data and ensuring compliance with industry standards.
3. Role-Based Access Control (RBAC) is essential for maintaining a secure environment by limiting access to authorized personnel.
4. The choice between cloud storage and on-premises storage requires a careful assessment of an organization's specific security and operational needs.
5. Data encryption, both at rest and in transit, is a critical component of data protection, ensuring the confidentiality and integrity of sensitive information.
6. In the face of evolving ransomware threats, practitioners and providers must employ robust security programs that can adapt to changing adversarial tactics, including regular backups, employee training, network segmentation, patch management, and the use of advanced endpoint detection and response solutions.

Chapter 16

Cross-Border Discovery: A Guide to Practical Challenges for US Counsel

About this Chapter

In this chapter, we will discuss practical guidance for US counsel navigating the logistical and operational challenges that arise in a typical cross-border matter requiring information from a foreign jurisdiction to be imported into the US for purposes of discovery.

16.1 INTRODUCTION

Economic globalization continues to drive the movement of data around the world within and between multinational organizations. As a result, lawyers in the US handling litigation and regulatory matters for those organizations are seeing a steady increase in activity that crosses international borders. For example, it's quite common for a commercial dispute in a US court to involve witnesses and information located in another country. Similarly, investigations into potential corporate wrongdoing—under the Foreign Corrupt Practices Act, for example—often implicate employees in multiple countries across different continents.

The legal issues surrounding the cross-border movement of data and information are exceedingly complicated. Privacy regulations, blocking statutes, national data security laws, banking secrecy laws, and other international regulatory regimes place a variety of restrictions on the movement of data from foreign jurisdictions into the US for legal matters. Also, different approaches to attorney-client privilege and the protection of trade secrets and other propriety information can impact counsel's strategic decisions. US counsel should engage with in-country counsel who have expertise with these laws and experience dealing with cross-border legal matters.

The aim of this chapter is not to untangle that knot of complex legal issues. Instead, its purpose is to provide US counsel with practical guidance in navigating the logistical and operational challenges that arise in a typical cross-border matter requiring information from a foreign jurisdiction to be imported into the US for purposes of discovery.

16.1.1 Social and Cultural Challenges

Some of the challenges likely to be encountered involve cultural differences and divergent social and workplace norms in countries outside the US.

First, local personnel who are asked to help gather information may be unfamiliar with the discovery process in the US and its requirements to produce extensive information. US-style discovery—which can be broad, burdensome, and intrusive—simply doesn't exist in many foreign jurisdictions. Unless the custodian is experienced or familiar with the process, they might be resistant to carrying out instructions from counsel. Education and collaborative communication are often necessary to secure their cooperation.

Language barriers can also pose challenges to effective coordination and project management. Even employees with solid English-language skills can be uncomfortable dealing with legal matters and talking to lawyers in a non-native language. The best practice is to use bilingual personnel who can translate as necessary and help facilitate the process.

If the information collected contains foreign language content, it is critical that the professionals working with the documents are proficient in the language. If search terms or analytic techniques are being used to search and cull the data, people comfortable in both English and the foreign language should be engaged so that they can effectively translate the search criteria while taking into account the nuances of the foreign language. Likewise, document review teams must be staffed with professionals who can demonstrate their facility with the foreign language, and who may need to be licensed in the local jurisdiction. If documents will be presented to a court or regulator, a certified translator should be retained to prepare the translated documents.

In some countries, there may be additional stakeholders in the information collection process that don't exist in the US, such as Works Councils. Works Councils, which are common throughout

Europe, represent workers at the local level and are similar to a US labor union. If there is a Works Council in place, you may need to consult with the Council or allow them to be involved before you interview an employee or collect information in their possession. There may also be circumstances where you need to notify a local data protection authority or other regulator before embarking on a data collection from employees.

Finally, another significant cultural difference in many countries as compared to the US is the approach to the work week and holidays. Working "overtime," working outside standard workday hours, and working on weekends are much less common outside the US. Similarly, employees in other countries are more likely to take extended vacations with no expectation that they will be asked to check e-mail or do other work while on leave. US counsel need to take these different work expectations into account when planning custodian collections in foreign jurisdictions.

16.1.2 Technical Challenges

Another set of challenges that counsel might have to address in a cross-border matter involves more technical issues.

Some foreign languages, including certain Asian and Cyrillic script languages, use different alphabets and grammatical constructions, which can present technical issues when processing electronic documents using standard western tools and protocols. In some instances, the characters may be entirely pictographic, so the tools we commonly use in the US to process, search, analyze, and review information may not be able to capture these differences. The de facto standard for character encoding for western languages—Unicode—does not always accommodate Asian languages (often referred to as "CJK," for Chinese, Japanese and Korean). Each country has its own distinct code sets, some of which utilize multiple code sets, such as Japan. Also, some bespoke email programs used in Asia generate unusual file types that are unrecognizable by western tools. You may need to consider different tools, or possibly adjusting the settings of your tools, to accommodate these languages.

Also, systems and software used in other countries may be different than what US counsel and legal technologists are familiar with. It is important to identify the name and versions of software being used so that the appropriate forensic tool can be used to execute a data collection. Similarly, different communication platforms are often used outside the US—WeChat in China, Viber in Europe, and imo in Qatar, for example—which require expertise in conducting data collections from these systems. It's advisable to engage with in-country experts, such as the client's Information Technology staff or local forensic experts, who are familiar with these systems. These experts can ensure that you are accurately, completely, and defensibly collecting the information needed, and complying with the applicable discovery specifications and ESI protocols in the US jurisdiction.

Once a data collection is exported to the US and stored on local systems, counsel should consider whether the technical information security protocols in place where the data are hosted

are sufficient to satisfy the requirements of the exporting country. For instance, if data are coming from the EU, counsel will need to ensure that any provider of eDiscovery services in the US offers adequate evidence of its "Technical and Organisational Measures" to protect personal data as required by the EU's General Data Protection Regulation. To effectively assess the information security of imported data, you may need to engage with InfoSec experts.

16.1.3 Timing Challenges

As counsel experienced with cross-border discovery can attest, the process is almost always slower than domestic discovery efforts. US lawyers need to take this into account when building out realistic timelines for discovery projects, especially when negotiating agreements with opposing counsel or securing scheduling orders from the court or regulator.

One of the most significant factors contributing to time delays—not to be underestimated—is navigating different time zones. Consider an "urgent" request sent by counsel in California at 11:00 am pacific time on Friday to the litigation team in Europe. That request likely won't be received and acted on until Monday morning local time, which can be frustrating to counsel accustomed to near-immediate responses from her US-based team. Also, keep in mind time zones when running up against deadlines—a document production delivered electronically at 5:00 pm in Mexico City will miss a 5:00 pm deadline in Washington, DC.

Also, the slower pace of work and more regularized work hours in some foreign countries can contribute to discovery delays. In the US, a litigation support professional is expected to work late, or on a weekend, if necessary to meet a discovery deadline; the same cannot necessarily be expected of team members in overseas locations.

Finally, if physical storage media containing electronic information—or other tangible items of evidence—need to be sent to the US, the process of overseas shipping and clearing US customs can introduce significant delays.

16.1.4 Cost Challenges

Not only is cross-border discovery slower than in the US, but typically it's more expensive. Counsel should set expectations appropriately when preparing discovery budgets for clients in these matters.

The foreign resources required to carry out discovery processes are generally more expensive than domestic resources in the US. This holds true for the labor costs of personnel as well as technology fees. One of the factors driving up prices is the relative scarcity of these resources—because many foreign jurisdictions don't have a US-style discovery process, there are fewer professionals and service providers experienced in this area. Also, the volatility of exchange rates can contribute to wide, and unpredictable swings in the cost of a project.

16.1.5 Conclusion

US counsel handling matters for large multinational organizations are almost certain to encounter the need for cross-border discovery. And, as the world's economies continue to reach across international borders, even counsel representing small and mid-size companies will more frequently face the prospect of gathering information found in other countries. At the same time that counsel are expected to come up to speed on the thorny legal issues raised, they must

also be prepared to deal with the very practical, logistical challenges discussed above. But the challenges are not insurmountable. By engaging with experienced, knowledgeable professionals in the local jurisdictions, and by preparing ahead of time for the inevitable process complications, US counsel can successfully navigate these difficult—but oftentimes interesting and exciting—international matters.

16.2 KEY TAKEAWAYS

There are five key takeaways from this chapter to remember:

1. Some of the challenges likely to be encountered in cross-border discovery involve cultural differences and divergent social and workplace norms, including unfamiliarity with US-style discovery, language barriers, additional stakeholders, and the general approach to work hours and holidays.
2. Another set of challenges that counsel might have to address in a cross-border matter involves more technical issues, such as processing and display of languages using non-Latin alphabets or collection from non-Western systems and software.
3. As counsel experienced with cross-border discovery can attest, the process is almost always slower than domestic discovery efforts, and US lawyers need to take this into account when building out realistic timelines for discovery projects, especially when negotiating agreements with opposing counsel or securing scheduling orders from the court or regulator.
4. Not only is cross-border discovery slower than in the US, but typically it's more expensive. Counsel should set expectations appropriately when preparing discovery budgets for clients in these matters.
5. By engaging with experienced, knowledgeable professionals in the local jurisdictions, and by preparing ahead of time for the inevitable process complications, US counsel can successfully navigate cross-border discovery.

Index

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

Consilio
Advanced
Learning Institute