

# How to Streamline and Execute Cost-Effective Investigations Across Disparate Data Sources

## TransPerfect Legal Solutions

Today's fast-moving and competitive business-to-business and business-to-consumer environment is driving organizations to engage in a variety of digital transformation processes, often contributing to the proliferation of data sources.

Indeed, the information technology infrastructure of businesses today is no longer limited to behind-the-firewall servers and enterprise appliances from traditional providers like Microsoft and IBM. Modern employees demand the ability to access their workspace and collaborate with colleagues in dynamic, diverse environments from the road, mobile devices, and the cloud.

A dizzying array of platforms have risen to meet this demand. The result is that present-day employees discuss projects on one platform (Microsoft Teams, Slack, Skype, etc.), collaborate through another (Confluence, SharePoint), and manage client relationships through another (Salesforce, Zendesk, Jira). And that's not to mention the underlying universe of corporate emails, text

messages, social media accounts, and accounting/finance databases. This panoply of data sources is a boon to operational efficiency, employee satisfaction, and client engagement, but it also a nightmare when it comes to legal, data privacy, and compliance professionals. Fortunately, tech-forward platforms and workflows have also taken shape to alleviate these pain points.

In this brief, we explore the tools and techniques for streamlined and cost-effective investigations across disparate data sources. From knowledge integration platforms to artificial intelligence and machine learning, we discuss how companies can access, search, and manage their enterprise data while remaining efficient, defensible, and compliant.



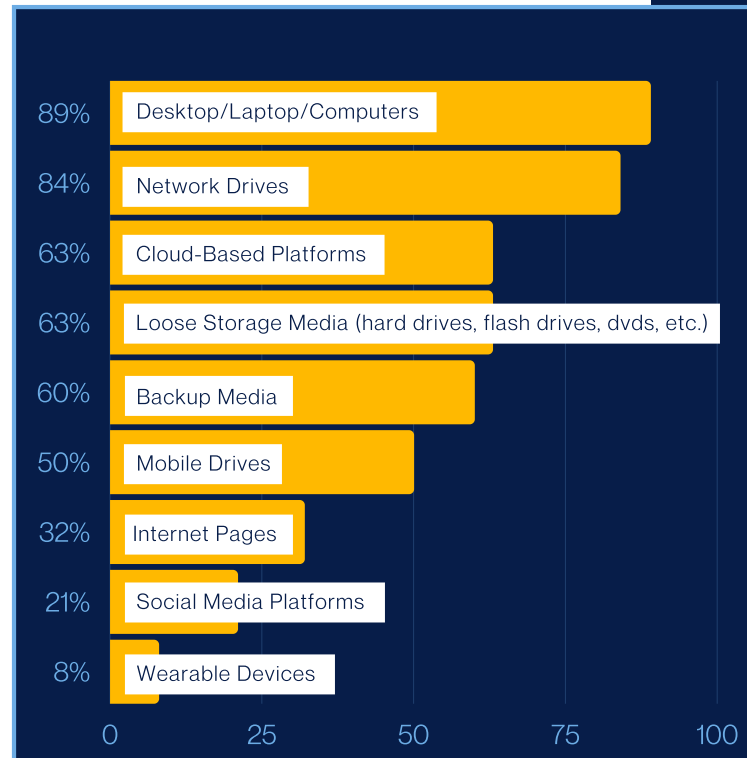
## Disparate Data Sources: A Snapshot of Modern IT Infrastructure and Challenges

One of the more impactful changes to the investigatory landscape has been the acceleration of new data source adoption in corporations and small businesses alike. Reliance on traditional sources for data storage and communication have changed; customary email and network share storage are being supplemented or undergoing outright replacement by cloud-based platforms, many of which offer a multitude of functions and serve in a hybrid capacity.

The utilization of additional data platforms has impacted the investigatory process at the point of collection in two ways; the need to preserve data from more platforms has grown exponentially and the format of how that data can be preserved has developed into a very real challenge. But first, we must discuss the impact of a seemingly endless variety of data sources.

While email and electronic documents still dominate data preservation, these practices have extended to databases, instant messages, text messages, multimedia files, and social media. Additionally, the breadth of devices capable of storing this data has expanded, and so have the devices companies are choosing for preservation. Most notable is the rise of cloud-based platforms and mobile devices, both fairly new to the investigation process and both presenting unique challenges. Aside from these sources, email, social media, and instant messaging are impacting investigations.

The above graphic includes data from a 2017 benchmarking report of in-house lawyers and demonstrates the breadth of platforms that fall within the scope of a modern litigation or investigation.(a)



## Email

The preservation and retention of email has changed significantly over the course of the last few years, due in most part to Microsoft Office 365's Security & Compliance platform and the Google Workspace of cloud-based business tools, which have continued to evolve into full-fledged investigation and analysis platforms. Importantly, as it relates to email collection for investigation purposes, these platforms reduce the need for onsite, in-person collections, and provide a universal mechanism for dealing with the preservation of email and certain types of unstructured data.

These developments have eased the need for IT departments to spend a great deal of time learning how to export mail, apply legal holds, and run targeted searches. In turn, the rise of companies "self-collecting" data has significantly streamlined investigations. According to a 2016 Gartner survey, 78% of enterprises used or planned to use Microsoft Office 365, which was a 14% increase since 2014.(b)

(a) - [www.bdo.com/insights/business-financial-advisory/2017-in-house-legal-benchmarking-report](http://www.bdo.com/insights/business-financial-advisory/2017-in-house-legal-benchmarking-report)  
(b) - <https://blog.x1discovery.com/2017/05/25/microsoft-office-365-is-disrupting-the-ediscovery-industry-in-a-major-and-permanent-fashion/>

## Instant Messaging

Instant messaging has been around for quite some time, but utilization in corporate environments has changed, along with the need, and ability, to preserve that data. Several factors, including speed, team collaboration, and mobile devices, have pushed the popularity of IM applications. Additionally, instant messaging applications are available across a wide variety of devices; the same application can be used on a computer, mobile device, tablet, or web browser to name a few. Certain instant messaging applications may offer a retention option, but in large part, the need to collect that data directly from the end-device still dominates collection requests. This can be particularly challenging when the messaging application is primarily used on a mobile device.

## Mobile Devices

The need to collect data from mobile devices continues to increase, undoubtedly impacted by the rise of “bring your own device” (BYOD) or “company-owned, personally-enabled” (COPE) corporate mobile device policies. In fact, a new abbreviation has emerged—Bring Your Own App (BYOA)—that really speaks more about the impact mobile devices have had on data collections. A recent article estimated that over 90% of knowledge workers use third-party applications for work.<sup>(c)</sup> Many of these applications are not officially approved by IT departments; they are applications the user has unilaterally adopted and put in use, a practice referred to as “shadow IT.”

This greatly impacts data collections because retention by the user is not happening often and the mobile device is the main source for preserving this data. Mobile device collections significantly differ from other digital devices, and application support is highly dependent on the make, model, and operating system. Furthermore, security plays a prominent role in the success of the data collection. Although recent advances to bypass/crack passwords have occurred, the ability to do so often

requires expensive mobile forensic tools or tools that are only available to law enforcement or the military. In short, apart from the device itself, in most instances the password is required for successful data collection.

Application support varies depending on the device. An application like WhatsApp may be supported for extraction on a certain mobile device, but not for another. Additionally, certain applications may not store the data, or a good portion of the data, on the mobile device but rather in the cloud. A recent trend is the evolution of mobile device forensic tools to have the ability to “reach” cloud accounts authenticated on the mobile device to collect data. Undoubtedly, this trend will continue in the future.

## Social Media

Social media platforms continue to evolve in terms of enabling users to collect and preserve their data. In recent years, tools that would crawl social media sites were the industry preference. However, due to recent events and privacy concerns, sites like Facebook and Instagram (owned by Facebook) have limited this ability of third-party platforms, forcing users to look for alternative options. The majority of sites have a native application feature that exports site data. The format of data from the archive option is typically the biggest differentiator, but it offers an easy, and cheap, option for data collection.

Since social media sites offer various methods of communication, depending on the site, data collection may occur from a mobile device or computer synced to the account.

## Cloud-Based Platforms

The popularity of project management and/or project collaboration platforms has had a significant impact on data collection and preservation. These platforms usually are cloud-based, offer several different utilities for project

(c) - <https://www.gartner.com/smarterwithgartnerbring-your-own-app-strategies/>



management and communication, and present unique preservation issues.

Platforms like Slack, Jira, and Quip provide the user with many operational benefits, including the ability to create projects, edit documents, and communicate via instant messaging. This is very efficient for the user but presents challenges for data collection since vastly different file types will be present within one platform.

Additionally, platforms like these primarily run from cloud servers, so the timing to export large volumes of data needs to be considered. In many instances, native export options provide a portion of the data, but may not include attachments, which need to be exported separately. The format of data stored on cloud platforms also needs to be considered. Data from cloud-based platforms usually has limited native options for export, especially when the need is a bulk export. If data is needed for legal review, the export format may not be suitable, and further processing of the data will be required to allow for human review.

In conclusion, the significant proliferation of data sources in recent years has resulted in numerous challenges to an efficient document preservation, collection, and investigation process. These challenges not only undermine the efficiency of an investigation but, as discussed in the next section, also create an obstacle to meeting a company's document retention and preservation obligations.



## The Legal Obligations: How to Retain and Preserve Electronic Documents and Data

The duty to retain and/or preserve electronically stored information (ESI) arises in a number of contexts. The most common sources of these obligations are document retention requirements imposed by statute and document preservation obligations triggered when ESI is potentially relevant to pending or anticipated litigation, arbitration, or governmental investigation. This section will survey such obligations under the laws of the U.S. and the U.K.

### Regulatory Requirements for Retaining ESI

#### 1. *The United States*

In the United States, obligations to retain documents and ESI for pre-determined periods of time come from a panoply of federal, state, and municipal laws. For example, under U.S. federal law, employers must retain records regarding employee benefit plans for six years. *See 29 U.S.C. S. 1027*. Likewise, under Section 17(a) of the Securities Exchange Act of 1934, a six-year record retention obligation is imposed on registered brokers and dealers with respect to a wide variety of documents (including all client account terms and all records needed to respond to an SEC audit).

The Internal Revenue Code contains a spectrum of potentially applicable retention periods depending on the action, expense, or event that the document records.(d) U.S. state insurance regulators similarly impose an array of retention obligations depending on the jurisdiction and record type. *See, e.g., 11 NYCRR § 243.2(b) (New York imposes a six-year retention period for policy records); 39 Pa.B. 4664 (Pennsylvania imposes a seven-year requirement); 114 CSR 15 (West Virginia imposes a five-year requirement).*

(d) - <https://www.irs.gov/businesses/small-businesses-self-employed/how-long-should-i-keep-records>

Document retention obligations are also imposed by industry-specific associations. As one example, the Rules of Professional Conduct regulating the practice of law contain a variety of obligatory retention periods depending on the relevant U.S. state. *See e.g., Colo. RPC S. 1.15-1.16 (imposing six or ten-year record retention provisions depending on record type); New Jersey Court R. 1:21-6 (seven-year retention period).* As another example, the American Petroleum Institute, a national trade association representing all facets of the oil and natural gas industry in the U.S., has promulgated a standardized Quality Manual that mandates that all member entities retain a diverse array of records for at least five years. Analogous provisions exist in industry standards governing accountants, automobile manufacturers, chemical companies, and insurance provider, among others.

## 2. The United Kingdom

Likewise, in England and Wales, obligations to retain documents and ESI can be found in all manner of statutes, regulations, and directives. While the examples below are by no means exhaustive, their diversity indicates the depth to which document retention is now a function of the modern legal and commercial landscape.

It is well known that companies must retain a copy of the minutes and resolutions from board meetings from the date of the meeting for 10 years—section 248, Companies Act 2006. Failure to comply renders every officer of the company guilty and liable to summary conviction. The same employer is also under an obligation to retain maternity pay records for three years after the end of the tax year in which the maternity pay period ends—regulation 26, Statutory Maternity Pay (General) Regulations 1986 (SI 1986/1960).

Unsurprisingly, there are a raft of obligations relating to medical and safety records. Schedule 3 of the Control of Substances Hazardous to Health Regulations 2002/2677 obliges employers to retain a list of employees exposed to substances

that can cause human disease, indicating the type of work carried out, the agent to which they were exposed, and records of accidents and incidents for a minimum of 40 years.

The Solicitors Regulation Authority Handbook sets out the rules applicable to law firms in England and Wales. As in the U.S. it stipulates at Rule 10 that certain records made under the SRA rules (including but not limited to instructions, transactions, commissions, etc.) must be retained for at least six years. In addition to the typical rules for companies, medical records, and professional organizations, records must also be retained for everything from environmental purposes to general tax inspection.

## 3. Document Retention Policies and Schedules

Because the duty to retain documents and ESI is derived from a complex web of statutes and industry standards, many companies invest significant time and resource development into maintaining and periodically updating document retention schedules and policies.

A document retention schedule can be thought of as the “what” behind the company’s retention obligations. This is typically a lengthy matrix that breaks down the entity’s records by category (e.g., accounting, legal, corporate, human resources), with each category being further broken down by record class (e.g., tax documents, contracts, employment applications, etc.). The retention schedule will detail the applicable retention period and cite the legal source of the relevant obligation.

A document retention policy can be thought of as the “how” and “who” of the company’s retention obligations. This document explains to employees and corporate stakeholders the scope of their retention obligations and how to comply with them.

The recent proliferation of document *destruction* obligations is further complicating matters—i.e., statutory or industry-based obligations *not* to retain data longer than necessary. As examples, under the GDPR, documents containing personal data shall be retained “no longer than is necessary for the purposes for which the personal data is processed[.]” *See GDPR Art. 5(1)(e)*. Likewise, the Payment Card Industry Data Security Standard (PCI DSS) obligates covered entities to “limit data storage amount and retention time to that which is required for legal, regulatory, and/or business requirements.” *See PCI DSS Requirement 3.1*.

In conclusion, a company’s document retention obligations are varied and multifaceted. They depend on the jurisdiction(s) in which the company operates, its industry, and its professional affiliations. The process of identifying applicable obligations is itself a challenge. Effectively implementing those requirements across a disparate matrix of data sources raises the challenge by several orders of magnitude.

## The Duty to Preserve ESI

The statutory and industry-based document retention obligations described above are constant requirements triggered by jurisdiction and verticals, not by specific events or circumstances that arise and disappear over time. By contrast, the document preservation obligations that apply when an entity is subject to a litigation, arbitration, or governmental investigation are ephemeral.

### 1. The United States

Dispute-based preservation obligations under U.S. law have been aptly summarized by the Sedona Conference as follows: (e) “whenever litigation is reasonably anticipated, threatened, or pending against an organization, that organization has a duty to undertake reasonable and actions in good faith to preserve relevant and discoverable information and tangible

evidence.” *See The Sedona Conference® Commentary On Legal Holds: The Trigger & The Process, 11 Sedona Conf. J. 265, 267 (2010) (“Sedona Legal Hold Commentary”)*. Thus, the duty to preserve documents can arise even before a lawsuit is filed so long as a party is on notice that future litigation is likely. *See Cache La Poudre Feeds, LLC v. Land O’Lakes, Inc., 244 F.R.D. 614, 621 (D. Colo. 2007)*.

Once a party’s duty to preserve documents has been triggered, the party is obligated to take comprehensive and multifaceted measures. *See Voom HD Holdings LLC v. Echostar Satellite LLC, 93 A.D.3d 33, 41-42 (1st Dep’t 2012)*: the party must (i) “take active steps to halt” any “automatic deletion features that periodically purge electronic documents such as emails,” (ii) “direct appropriate employees to preserve all relevant records,” and (iii) “create a mechanism for collecting the preserved records so that they might be searched by someone other than the employee.” *See also Sedona Legal Hold Commentary at 267*: “the duty to preserve requires a party to identify, locate, and maintain information and tangible evidence that is relevant to specific and identifiable litigation.”

Implementing such measures across a disparate array of data sources requires a deep understanding of how those data sources operate, and a potentially significant commitment of resources. *See generally “The Sedona Principles, Second Edition: Best Practices, Recommendations & Principles for Addressing Electronic Document Production 17” (The Sedona Conference Working Group Series, 2007)*: “Transaction costs due to electronic discovery” can be “overwhelming.”; *Concord Boat Corp. v. Brunswick Corp., No. LR-C-95-781, 1997 WL 33352759, at \*4 (E.D. Ark. Aug 29, 1997)*: “Hard disk or tape storage of data is very costly. With

(e) The Sedona Conference is one of the leading think tanks on electronic discovery issues, whose members and authors consist of U.S. judges, private practitioners, and legal academics.

corporations spending enormous amounts of money to preserve business-related and financial data . . . they should not be required to preserve every e-mail message at a significant additional expense.”

In recent years, parties to litigation have struggled to satisfy their document preservation burdens with respect to the more technically progressive and innovative data platforms discussed above, resulting in sanctions, fines, and reputational harm.

For example, in *Brown v. Tellermate Holdings, No. 2:11-cv-1122, 2014 WL 2987051 (S.D. Oh. July 1, 2014)*, the court imposed severe sanctions against a litigant for failing to properly preserve and produce ESI records stored in the company’s Salesforce.com account. In that litigation, as part of their employment discrimination claim, plaintiffs demanded that the defendant produce reports from Salesforce, a cloud platform that Tellermate used to track employee sales performance. The defendant refused, arguing that it could “only access the salesforce.com database in real time,” and thus, if plaintiffs desire historical data, they would need to subpoena Salesforce itself. See 2014 WL 2987051, at \*1, \*5, \*20. In reality, however, this argument betrayed a fundamental misunderstanding of the Salesforce platform; “any [defendant] employee with a login name and password could access . . . historical information [on salesforce.com] at any time.” *Id.* at 9. The court noted that the defendant’s “failure to appreciate” the nature of the defendant’s ESI led to a “corresponding failure to take steps to preserve that information” beyond the three- to six-month period. it was automatically stored by salesforce.com and as a result, relevant data was lost forever. *Id.* at \*9.

As a result, the court imposed a variety of sanctions against the defendant, including a preclusionary order prohibiting the defendant from introducing any evidence for performance-related termination of the plaintiffs, effectively eviscerating the defendant’s core defense in the litigation. Notably, the Tellermate

court did not only sanction the litigant for its failure to preserve ESI stored on Salesforce, but also outside counsel. See 2014 WL 2987051, at \*1: chastising counsel for falling “far short of their obligation to *examine critically* the information which Tellermate gave them [about ESI].”

## 2. The United Kingdom

Likewise, in England and Wales, there exist numerous obligations around ESI and sanctions for non-compliance. The most common penalties still come in the form of costs. Parties and their representatives are well aware of the risks and must advise their clients accordingly. For example, in *West African Pipeline Company Ltd v. Willbros Global Holdings Inc (2012) EWHC 396 (TCC)*, court-imposed cost sanctions were applicable to seven separate breaches ranging from failure to properly gather custodians’ data to failure to provide appropriate metadata fields.

The specific obligation to preserve documents in the context of a dispute, emanates from *Part 7 of Practice Direction 31B—Disclosure of Electronic Documents in the Civil Procedure Rule*: “as soon as litigation is contemplated, the parties’ legal representatives must notify their clients of the need to preserve disclosable documents. The documents to be preserved include Electronic Documents which would otherwise be deleted in accordance with a document retention policy or otherwise deleted in the ordinary course of business.” This rule, however, is subject to amendment. At the time of publication, new disclosure rules for the Courts of England and Wales are being considered by the Rolls Building Disclosure Working Group in response to widespread industry concern around the scale and complexity of disclosure. The resulting rule changes will no doubt examine the growing number of disparate data sources in which information is contained and how best to deal with them procedurally.



While commonly accepted in progressive jurisdictions, it is worth noting that the definition of a document under the CPR is so broad—“anything in which information of any description is recorded”—that any format of ESI is a document in both rules and case law.

It is well established that the destruction or failure to preserve such documents would draw adverse inferences from a court, potentially harming their credibility and the veracity of a case. As mentioned above, financial penalties for breaches of such obligations are common. Where it appears documents have not been properly preserved, the court has further powers relating to recovering such information, for example to compel forensic recovery of deleted data. On a strict interpretation of CPR r.3.4 (2)(c), breach of such direction practice would provide grounds for strike-out. However, relevant jurisprudence suggests the court will only go this far if such destruction is an attempt to pervert the course of justice. *See Douglas v. Hello! (2003) EWHC 55 (Ch)*. In practice, therefore, lawyers are obliged under the procedural rules to notify clients of the need to preserve disclosable ESI from any sources where information relevant to a particular action may be stored. It is of course more important where organizations have routine procedures relating to any electronic information.

The starting point for establishing which sources fall into scope of disclosure is CPR r. 31.5. The current procedural rules require parties to state where and with whom electronic documents are stored (CPR r.31.5 (3)(b) & (c)) 14 days before the first Case Management Conference (CMC). Additionally, US style “meet and confer” obligations are foisted upon the parties before the CMC. This ensures, to the extent possible, no relevant information slips through the net at the earliest stage.

The courts’ appreciation for disparate data sources was clear in the recent case of *Glaxo Wellcome UK Limited & Anor v. Sandoz Limited & Ors (2018) EWHC 1626 (Ch)*. The claimants made an application

in relation to the defendants’ disclosure, regarding a particular use of a “DocXchange” platform. The application was made in the context of what was already considered to be significant disclosure. The claimants sought over 40 custodians, over time periods in excess of 10 years. Even though extensive search terms had been applied, over 400,000 documents were still manually reviewed. The process took six months and cost over £2 million.

The platform in question was set up for various defendants to share information in relation to the inhaler design in question. The issue was that the platform was destroyed when some of the defendants were in the process of joining the action. The defendants’ solicitors had provided, on affidavit, information about the platform and its destruction but the judge noted that it was lacking in detail and importantly came from a lawyer, not a technologist who had an appreciation for the information on the platform or reasons for its decommission. The judge said, “It is not clear from Mr. Howe Q.C.’s evidence, who represents the defendants and accepts the fact that there was a likelihood of documents being held in the platform which were not held elsewhere. There is no uncertainty about that. It follows that there may have been documents falling within CPR 31.6 within the platform which should have been disclosed during the destruction of the system.”

Furthermore, “There is no suggestion from the court that there has been an attempt to consciously mislead. However, the exercise of providing disclosure is underpinned by duties placed on the disclosing party to undertake the exercise with due care. The evidence that has been provided to the court suggests that the defendant did not exercise proper care in this case. The defendants and the defendants’ solicitors were plainly aware of their obligation to disclose documents which they had in their control but which they no longer have.”



The judge was in favor of the claimants and made an order on the terms sought. This case is an important reminder that parties must be informed of their duty to preserve documents as soon as litigation is contemplated. Lawyers must properly explore all potential sources of information where material evidence can exist.

## Investigations Across Disparate Data Sources

The foregoing demonstrates that while data sources are expanding and changing rapidly, the legal obligations to retain, preserve, and collect such ESI, in the context of litigations, arbitrations, and governmental investigations, remain constant. This juxtaposition creates many challenges and mandates new solutions.

While commonly accepted in progressive jurisdictions, it is worth noting that the definition of a document under the CPR is so broad—“anything in which information of any description is recorded”—that any format of ESI is a document in both rules and case law.

## New Sources, New Challenges

Traditional sources of ESI were often accessed in a similar way. Data from computers and servers were imaged using traditional forensic tools and the file content extracted, processed, and reviewed. Preservation methods and defensibility were well understood, and various file types with associated metadata were combined to tell a story. Even scanned paper documents fit well into this workflow.

When mobile devices started to be recognized as important sources of evidence, challenges were created by dozens of different cell phone manufacturers, each with a proprietary operating system for their devices. Years of consolidation and standardization in the mobile device industry has helped to solve some of these problems, resulting in

a few acceptable phone operating systems that are largely aligned with computer operating systems from a technical standpoint.

In today’s technical landscape, we are once again seeing the proliferation of proprietary operating environments, across multiple technologies and business segments. Many new sources of ESI are being introduced almost on a daily basis, and virtually all of them are custom built. Some of these new sources (e.g., cloud-based storage and tools, or distributed ledgers) have some similarities to traditional data sources, driven mainly by the need to maintain backward compatibility. Others (e.g., machine learning, artificial intelligence, and distributed processing) represent new concepts with no direct similarity with the traditional process of investigating ESI.

These new data sources sound familiar from last year’s science fiction and this year’s marketing campaigns. The buzz words and broad technical concepts include: cloud-based, blockchain, cryptocurrency, machine learning, artificial intelligence, Internet of Things (IoT), and big data.

The business processes that utilize these emerging technologies are showing up today across the investigation process. Evidence from IoT sensor nets or decision-making processes for self-driving cars are now part of data investigations and regulatory inquiries. Automated decision making driven by big data, and financial transactions involving cryptocurrency or distributed ledgers are at the center of some court cases. To one degree or another, each of these technical innovations brings challenges to the process of retaining, preserving, collecting, and investigating ESI. Additionally, each product or system in a given category will not introduce the same challenges in the same way as few standards exist with vendors in these technologies. However, we can categorize the challenges and outline typical

approaches that can be applied broadly when addressing new or novel data sources in the context of a litigation or investigation.

## Strategies to Acquire, Normalize, and Present ESI from Modern Data Sources

The difficulties in integrating new data sources across the investigation process can be grouped into three categories: acquisition, normalization, and presentation.

### 1. Acquisition

Traditional data sources have long-established methods for surveying, assessing, and collecting ESI. Non-traditional data sources often require different approaches to identify relevant data, determining volumes, targeting the required data, and preserving it externally.

In traditional acquisition and preservation, the gold standard was a bit-by-bit preservation or image of the source storage system. This is still commonplace when preserving data in laptops, desktops, and storage devices. Email communications have typically been collected into a known container format such as PST. Sources like mobile devices and large file servers are often collected logically, preserving the data and control file structures from the source.

Many new data sources cannot be accessed directly. Their data is typically created and accessed through a software application or some other process. Examples include proprietary chat or communications platforms, and IoT management platforms that control devices connected by the internet. There are usually a few strategic options for obtaining data:

- **Source Application.** In some cases, the application that was used to create the data will allow some or all of the data to be exported. The export format may be limited, and the export

may or may not include metadata or internal control data, which is useful to determine when and by whom the data was created. The data may also be interpreted or abstracted, as is often the case in IoT, big data, or AI systems. Some of these data sources may also have an administrative console through which certain data can be exported. Whether such exports are forensically defensible—i.e., capture relevant metadata fields without impacting them during the export process—should be assessed and validated on a platform-by-platform basis.

- **Third-Party Application.** For some cloud platforms, block chain standards, and other types of new data sources, there are third-party applications available to help preserve and collect data. Examples include social media platforms (Facebook, Instagram, etc.) and group collaboration tools (Slack, Jira, etc.). These data sources may allow for direct data export, but third-party tools allow more control over the content and format of the data retrieval. Once again, whether such exports are forensically defensible should be assessed and validated on a platform-by-platform basis.
- **API.** Many cloud-based data sources have an Application Programming Interface (API) designed into them. This feature allows platforms to communicate with each other by passing data back and forth. The data available through an API often exceeds that available via other methods, though it tends to be raw and unformatted. APIs can be simple or very complex, and some platforms have multiple APIs with varying capabilities. The third-party applications previously mentioned often use APIs to access data. While most have an API of some kind, the platform owner may not officially allow or support its use. In the absence of a third-party tool, using an API to acquire data may require custom programming or scripting.

- **Direct Data Access.** For some new data sources, the raw data may be stored in a database or file system. Gaining access can be problematic as these data sources are often multi-tenanted with no easy way to segregate data or access, causing the resource owner to object to this method.
- **Access, Security, and Debugging Logs.** Depending on the nature of the data source and the inquiry, sometimes information such as user access logs, security logs, or event and debugging logs may be relevant. These types of logs are often available for cloud-based data sources, though they may be transitory with fairly small retention periods.

## 2. Normalization

Managing the combination of data from novel sources and non-standard formats into a body of conventional ESI (e.g., email communications and business documents) can be difficult. Metadata, which often plays a role in understanding digital evidence, can have different meanings. Extracting new data in this context can make its meaning and relationships to other data unclear. The use of search terms and other data reduction techniques may need to be adjusted to accommodate the new data sources.

Indeed, when data from multiple sources is combined in a single investigation for a consolidated view of all available information, there may be certain aspects pertaining to it that need to be adjusted or synchronized to fit with the whole. What is required for each data source should be evaluated separately based on the needs of the case and the role that is expected to play. Areas for consideration include:

- **Metadata.** The taxonomy and meaning of metadata fields between data sources should be normalized. Different sources may treat fields like MAC dates and owners separately. Time zones should be synchronized. User or custodian names may be represented differently across sources and require synchronization.

- **Threading.** In modern technology environments, it is not unusual for a conversation about a single topic to take place across multiple platforms. If those conversations need to be threaded together, the user names, data, and conversation members will need to be synchronized for continuity.
- **Enrichment.** If logs or raw data are collected, the data may need to be enriched for it to be properly synchronized. IP addresses may need to be linked with names of people or companies. Internal reference numbers may need to be correlated with other data from a raw data dump. Logs may need to be pre-filtered to include only log entries of interest.
- **Volume.** If an extremely large amount of volume is collected, as may be the case when handling big data, IoT, or AI/MT systems, some pre-analysis may be necessary. The individual data points may be too voluminous to tell a story, but some basic numeric analytic or summaries (supported by the details) may better meet the needs of the case.

It is also important to understand if the data from a source has been pre-filtered or limited in some way. The purpose of normalization is to make sure the data tells a story in a consistent voice, and to cut out any differences that may impact downstream interpretation or analysis.

## 3. Presentation

In the investigation process, the last step often includes “producing” or “presenting” relevant ESI to an adversary, a governmental agency, or an internal compliance committee or board of directors. It may be difficult to format or present information obtained from new technologies. Sometimes there is no mechanism for displaying this data outside of the device or process that created it.

Once the data has been collected and normalized, there needs to be a coherent method for presenting it. Materials collected from traditional sources such as electronic documents, emails, or text messages, can be displayed in TIFF or PDF format, or native documents or text files. There are no standard formats for displaying or presenting large amounts of data from IoT sensors, or the decision tree used by an artificial intelligence engines to decide on an action. The review, presentation, and production processes should be adjusted to work within the limitation imposed by the nature of the data as well as the methods used to acquire and normalize the data. For example, IoT sensor data may be presented as a numerical and statistical analysis of the body of the data, accompanied by samples of the data itself. The presentation method needs to avoid interpretation of meaning and focus while presenting the data such that it can be understood in the context of the case.

One emerging solution to the integration of disparate data sources is the knowledge integration platform. These tools, generally a cloud-based service offering, can connect directly to multiple data sources simultaneously. Such platforms often use APIs to access the ESI in each data source, either as an on-demand function or continuously. As the platform acquires and aggregates the data, it

also addresses the issues of normalization and presentation. Some of these knowledge integration platforms add functionality such as automatic indexing, sophisticated search, and the use of AI to identify the file content, perform sentiment analysis, and identify languages present. While these solutions can be extremely useful in managing multi-sourced data collections, they are subject to inherent limitations of the platform or the APIs in use.

## Conclusion

There is no end in sight for the introduction of new and diverse data sources into the investigation process. While these technologies mature and evolve at an incredible pace, the underlying legal and regulatory duties to retain, preserve, collect, and analyze change at a glacial pace. So, while standards may start to emerge that ease the issues of acquisition, normalization, and presentation, the practitioner needs to be prepared to move quickly and understand the available options. As source diversification continues, practitioners will continue to refine the guidelines and methods required to incorporate this data into the investigation lifecycle.

### TLS Authors:

Al-Karim Markhani, Esq.  
*Vice President, Consulting*

Daniel S. Meyers, Esq.  
*President, Consulting & Information Governance*

