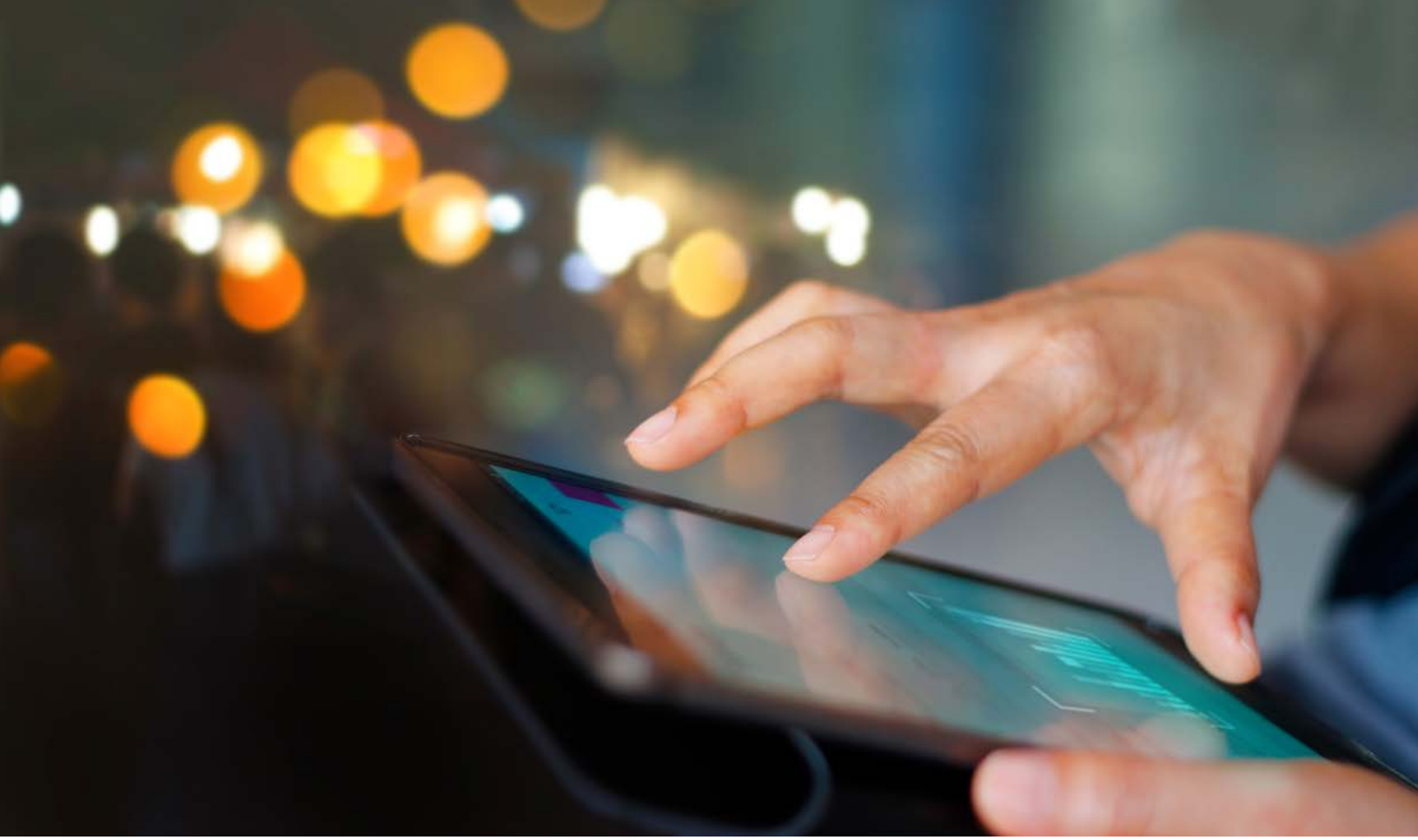
White Paper

epiq clarity

Supporting Nontraditional Data Types in eDiscovery

People. Partnership. Performance.



overview

It is no secret that the age of “big data” is upon us. We are experiencing an exponential growth in electronically stored information. In eDiscovery today, the path corporate data takes – from generation to collection to review and production - is all performed electronically. The important data we need is often swimming in a sea of millions of database records, located on the servers of some other company, or in the pockets of hundreds of employees. While data sharing and collaboration become easier, ownership of this data is becoming harder and harder to discern. Devices that were once self-contained and lacked the ability to store, transmit and share anything but the smallest amounts of data now easily store billions of bytes of data, and are connected to networks where that data can be shared, lost, compromised, or overlooked.

This white paper discusses how some of the newer and more difficult types of data are stored, identified, preserved, collected, searched, reduced, reviewed, and used in litigation. It is important to provide a basic framework for understanding how this data is handled today, and how sometimes that may not meet client/ corporate needs.

What is nontraditional data?

Traditional data typically refers to data that is user-created, organized, and has established workflows in eDiscovery. Items such as email, Microsoft Office files, or paper documents scanned to images and then OCR'd might all be examples of traditional data. Nontraditional data (in the context of this paper) is everything else, including but not limited to:

- Web and social media sites (Facebook®, Snapchat®, Instagram®, Twitter®)
- Text/Instant messages
- Chat and Group/Collaboration applications (i.e. WeChat®, WhatsApp®, Yammer®, Slack®, MSTeams®, Trello®)
- Audio recordings
- Internet of Things (IoT) devices such as FitBits®, WiFi connected Thermostats, Amazon Echo®, etc.
- GPS data, vehicle telemetry data
- Other Structured Data (SAP® and other financial systems, SQL databases)
- Data saved to cloud platforms (Salesforce, Dropbox, Google Drive)

These data types are becoming more prevalent as companies leverage information such as customer behavior, spending trends, driving habits of insurance customers, trucking fleet GPS data, and the like to streamline their business operations.

In dealing with nontraditional data, there are several challenges to be addressed. Companies may not be aware of the prevalence of nontraditional data within their enterprise. Even where there is awareness, the data may exist in locations where it can be difficult to retrieve when needed. Within these new nontraditional data types there have also been exponentially greater volumes of data. For example, in the same time it might take a person to write one email, a GPS logger or similar device might record hundreds or thousands of data points. Lastly, the context that is easily discerned from traditional data types may not be present for nontraditional data types, and may require additional analysis to determine what the data means.

What is discoverable?

Nontraditional data is discoverable under the same rules as traditional evidence. As defined in Federal Rule of Civil Procedure 26(b), the scope of permissible discovery is:

> Parties may obtain discovery regarding any non-privileged matter that is relevant to any party's claim or defense and proportional to the needs of the case . . . . Information within this scope of discovery need not be admissible in evidence to be discoverable.

Until recently, if such data was relevant, it has been discoverable unless there was an exception (e.g. undue prejudice, unduly burdensome to produce, etc.) provided for in the Federal Rules of Evidence (FRE) that permitted a party to not disclose it.

The relevance requirement has traditionally been an extremely low bar to clear under the FRE. Federal Rule of Evidence 401 establishes 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." However, recent changes to the FRCP (2015 amendments) attempt to reduce the overly broad production of data by implicitly including the concept of proportionality, and we are beginning to see emerging case law that leverages this concept.

Case law supports the position that nontraditional data must be preserved and collected when relevant. In *Regas Christou v. Beatport, LLC*, 2013 WL 248058 (D. Colo. January 23, 2013), the court sanctioned the defendant for "taking no steps to preserve the text messages," leading to a spoliation sanction. In *Thurmond v. Bowman*, No. 14-CV-6465W (W.D.N.Y. Mar. 31, 2016) the defendants sought spoliation sanctions against the plaintiff over the disappearance of numerous allegedly relevant posts from her Facebook account. In this case the plaintiff made a privacy settings change to the account that changed the public visibility of the posts, but had not deleted them. The plaintiff eventually produced the materials and the court declined to apply sanctions, but the court did express its displeasure with the plaintiff's "troubling" conduct, and indicated that sanctions might be applied if there was any further inappropriate conduct. Failure to preserve or produce responsive content can have serious consequences. *United States v. Vaughn*, No. 14-23 (JLL), 2015 WL 6948577 (D.N.J. Nov. 10, 2015); *In re Pradaxa (Dabigatran Etexilate) Prods. Liab. Litig.*, MDL No. 2385, 2013 WL 6486921 (S.D. Ill. Dec. 9, 2013); *Southeastern Mechanical Services, Inc., Plaintiff, vs. Norman Brody, et al., Defendants*, 2009 U.S. Dist. LEXIS 85430. However, there are still limitations. For example, the court in *Shenwick v. Twitter*, N. 16-CV-05314, 2018 WL 833085 (N.D. Cal. Feb. 7, 2018) concluded that the defendant did not need to disclose direct messages produced by individual custodians who were not parties to the lawsuit.

Congress also passed the "Clarifying Lawful Overseas Use of Data" Act (CLOUD Act) in 2018 that extended the reach of subpoenaed material internationally. The CLOUD Act, which amended the Stored Communications Act of 1986 (18 U.S.C. § 2713), allows parties to demand (via subpoena or warrant) information from American technology companies even if the data is located outside of the U.S., which is a common occurrence in today's globalized digital atmosphere. While this law is a criminal statute that applies to criminal cases, civil litigators may see a ripple effect from this law and should be prepared to comply with demands for nontraditional data sources that are stored on foreign soil.

Identification and control issues

An initial challenge with preserving and collecting nontraditional data is dealing with the issues surrounding possession, custody, and control of that data. Often nontraditional data is not actually stored where people think it is stored, and is sometimes gathered and sent elsewhere. For instance, much of the data viewed on a mobile device may not actually be stored on that device, and is instead located in the cloud or on an enterprise server. To make things more complicated, in many instances discovery will involve a third-party that hosts the data. All the possession, custody, and control issues associated with such arrangements in traditional discovery will come into play when information is needed from third-party hosts. Some data may require a subpoena in order to obtain it, and still other data cannot be disclosed absent the issuance of a search warrant or similar court order.

As discussed above, the CLOUD Act addresses this issue in certain situations. The law states: "A provider of electronic communication service or remote computing service shall comply with the obligations of this chapter to preserve, backup, or disclose the contents of a wire or electronic communication and any record or other information pertaining to a customer or subscriber within such provider's possession, custody, or control, regardless of whether such communication, record, or other information is located within or outside of the United States." This greatly expands the definition of control and mandates that organizations holding data subject to this law preserve and disclose it to law enforcement even if it is not in the organization's physical possession and is housed internationally.

The CLOUD Act was a product of a case where the government was attempting to get data via warrant from a U.S. organization stored on a server located in Ireland that was pertinent to a criminal investigation. The organization refused compliance because it housed the data internationally. Obviously, the CLOUD Act made this argument moot and the Supreme Court vacated the case in *United States v. Microsoft Corp.*, No. 17-2, 584 U.S. ___ (2018). This significantly expands potential preservation and collection obligations beyond U.S. borders.

Civil case law also seems to be heading in this direction - that courts will hold organizations responsible for nontraditional data preservation

even when the organization does not have physical control over the data. The case of *Ronnie Van Zant, Inc. v. Pyle*, 2017 WL 3721777 (S.D.N.Y. Aug. 23, 2017) extends this obligation to independent contractor data. In this case, a consultant lost text messages when switching cell phone providers, which prohibited the organization from turning that data over during litigation. The court deemed the texts to be within the organization's control and awarded sanctions for failing to preserve the data.

This duty can also extend to nontraditional data housed by a third-party in the ordinary course of business. There generally needs to be evidence of a contract in order for the possession, custody and control obligations to kick in. One common example is Salesforce, which is a CRM cloud platform that many businesses use. It integrates all facets of a customer or client's profile into one shared view that all authorized employees can work in and view. *Williams v. Angie's List*, Inc., 2017 WL 1318419 (S.D. Ind. Apr. 10, 2017) was a case where individuals claimed their employer was underreporting the hours they worked and therefore underpaid them. The employees commonly used Salesforce for work tasks and requested background data from the site during discovery to help prove their case. The court deemed the Salesforce data within the organization's control because it had a contractual right to obtain the data and logging background data when employees worked was part of the ordinary course of business between Salesforce and the employer.

Some of these possession, custody, and control challenges can be mitigated through policy creation and prelitigation readiness. One of the tools that can assist in this process is a data map. This is the process by which the entire data inventory of an organization is mapped, and it identifies the applications generating the data, and where the data is ultimately stored. It should differentiate between convenience copies of data stored in multiple locations and data stored in systems of record. If such a document does not already exist, corporate disaster recovery and business continuity plans are often good places to start, as they generally identify many of the critical business applications.

Once a data map is completed, corporate policies should be evaluated to ensure that unsanctioned applications are restricted, and existing retention policies are universally applied to nonstandard data sources. Many of the applications that generate and store nontraditional data can be configured to comply with these corporate retention policies.

For nontraditional data types that are frequently used in litigation, specific and documented workflows can be created so that productions of this data can be performed in a consistent and repeatable fashion. This may involve working with third-party vendors of these applications to create export facilities that meet corporate needs.

Creating governance policies, data maps, and custom workflows are crucial and provides a foundation for properly addressing nontraditional data. Failing to take these steps may result in litigation delays and negative outcomes, and some courts have even gone as far as to sanction organizations for not having a mandatory ESI preservation policy. In *Sell v. Country Life Ins. Co.*, 189 F.Supp.3d 925 (2016), the court held that only requiring employees to preserve emails after there is a litigation hold in place was deficient because had this policy applied at all times, the party would have preserved emails that were crucial to the claim.

Collection issues

Data collection can be a challenge for any company, and the challenges become magnified when dealing with nontraditional data sources. The collection of nontraditional types of data often results in issues when the same tools and workflows used to collect traditional data are employed. These methods may not work very well, or at all, and more advanced methods of data collection may be required in order to be successful.

A good example of these difficulties is in the area of mobile device data collection. While the collection of mobile device data is becoming more and more common, the never-ending onslaught of new applications for mobile devices along with the frequent technological changes in the mobile device industry has made the collection of nontraditional data from these devices a significant challenge. Collection software may or may not support the collection, decryption, and parsing of data from new or recently updated devices and applications, and as these devices and applications increase their data security capabilities it can detrimentally affect the ability to collect and use relevant data. As an example, WeChat® messages can be collected and parsed from some phones running some operating systems with certain security patches, but not from others. Until collection tools build in support for the myriad mobile devices and applications, or simply catch up to changes that are made to these devices and applications, the only/best way to collect, display and produce this data may be to view it on the device and photograph each individual message or conversation. This is certainly not ideal.

Other types of nontraditional data may pose similar, yet different, collection issues. For instance, website collections can be accomplished in different ways depending on the needs of a particular litigation. Does counsel need to preserve a fully functioning offline replica of the website that can be used like an online version, or are hard copies of the web pages within a site needed for production to the court or opposing counsel? Many sites have elements such as scrolling text boxes and multimedia elements that cannot be easily captured and printed for review and production. If collected using a particular method, what happens if counsel later discovers that they cannot read critical text because it was in a scrolling text box and cannot be seen on a static copy? These factors must be considered, and new workflows created to address these issues, at least until the mainstream collection tools catch up and support these needs.

Structured data provides yet another example of collection difficulties. For companies whose important data resides in scalable transactional databases, the challenges can be significant. Such databases can have hundreds or even thousands of related tables containing relevant data, but the ability to collect it in ways that make sense and can be displayed for use in court may be limited to the export features of the application. These are sometimes limited, and the data often needs to undergo additional processing and/or conversion post-collection in order to be usable.

While the collection and preservation of nontraditional data is often necessary, one potential way to limit the difficulties of collecting it is for attorneys to argue proportionality under the recent changes (2015 amendments) to the Federal Rules of Civil Procedure. The argument that such data should not be considered because the likelihood of collecting and producing relevant data is outweighed by the likelihood of collecting and producing irrelevant data could be a successful one. Since enacting the amendments, courts have taken a much closer look at proportionality than ever before. However, parties must generally be able to clearly articulate a reason in favor of or against proportionality in order to be successful. What is particularly interesting since the enactment of the amendments is that some courts have gone a step further and looked at proportionality even where a party has not asserted it. See *Curtis v. Metro. Life Ins. Co.*, No. 3:15-cv-2328-B (N.D. Tex. May. 4, 2016). Here are some additional key proportionality cases:

Gordon v. T.G.R. Logistics, Inc. d/b/a T/G/R/ Transport; and Igor V. Varga, No. 16-CV-00238-NDF, 2017 WL 1947537 (D. Wyoming May 10, 2017) demonstrates such a successful argument. In Gordon, a TGR Logistics tractor trailer struck plaintiff Gordon's vehicle. Gordon sued, claiming that she suffered "numerous physical injuries, pain (back, neck and jaw), traumatic brain injury, post-traumatic stress disorder, anxiety and depression" as a result of the accident. TGR moved to compel the discovery of "an electronic copy of [Gordon's] entire Facebook account history." It was argued that Gordon's complete Facebook history — both before and after the accident — was relevant to its defense. Gordon refused to comply, countering that "the request is unduly burdensome, lacks relevance, and is overly invasive." The court observed that under FRCP 26(b)(1), discovery encompasses "any non-privileged matter that is relevant to any party's claim or defense and proportional to the needs of the case." (Underscore supplied). The court concluded that granting discovery for Gordon's entire Facebook account history "would provide minimal relevant information while exposing substantial irrelevant information." Therefore, it would "exceed the proper limits of proportionality."

In *Trainer v. Continental Carbonic Products, Inc.*, Case No, 16-cv-4335, 2018 WL 3014124 the court held that there was a lack of proportionality and declined to order production of certain text messages when the defendant requested several text messages that were already produced in a pre-litigation investigation. The judge predicated this decision on the belief that the texts would have minimal value to the defense's case.

In 2019, a federal court in Kansas further solidified the importance of proportionality under the 2015 amendments. The judge in *Russell v. Kiewit Corp.*, No 18-2144-KHV (D. Kan. June 4, 2019) held that an employee's request for his entire email personal storage file went against the intent of the amendments because such a broad request is not related to the tenets of the claim and did not have sufficient relevancy.

The CLOUD Act could also pose some potential nontraditional data collection problems down the road since it may be challenging to collect data from an international server. Conflicting laws and privacy obligations may present various roadblocks.

Data volume and data structure issues

The tools and technologies used in the processing and review of traditional data sources are quite mature and can get through significant volumes of data without too much difficulty. The toolkits available to process nontraditional data however, are lagging behind in their ability to keep up with the exponentially greater volumes of data being generated by modern applications and connected devices. This gap requires the eDiscovery practitioner to approach nontraditional data sets with a greater understanding of what each data set represents, and the potential matter specific questions that can be answered by the data.

The majority of nontraditional data sets when distilled to their simplest forms are structured databases. Messaging apps, collaboration tools and connected devices all utilize databases to capture and store information generated. The scale of these databases can vary significantly from large enterprise databases with thousands of tables to a mobile app that utilizes a much simpler database structure. Beyond the complexities of the structure itself, the sheer number of data points that may need to be analyzed can also present a challenge.

When dealing with databases with a complex underlying structure, it is important to obtain an understanding of how the database segregates and stores the various categories of information. Databases are usually designed for efficient storage in order to minimize the redundancy of information. For example, within the ordering system of an online store, customer details, product descriptions, and order details may be spread across several tables. When an order is placed, the system will generate an entry in the order table with a link to the customer information, and additional links to the product details. The customer information is only stored once, and the order table simply uses a single element such as a customer number to establish the relationship between the order and the customer. Likewise, the order table may use a single element such as a product number to establish a link between an order and the products included in the order. In this manner the customer details are stored only once in the database even when the customer makes hundreds of orders. Similarly, the product details are stored only once even if thousands of customers order the same product. A database design that uses this method to reduce redundant

information is referred to as a “normalized” database.

In the context of eDiscovery, a normalized database may not be ideal. Often there is a need to “flatten” this multi-table structure and create a simplified structure that is easier to work with. There may be a need to identify and isolate segments of information across multiple tables and carve out the relevant information into a new database or export file. A focused scope that limits the production to only the specific data elements needed to answer discovery questions will assist in reducing the complexity and volume of datasets that will need to be analyzed.

With regard to “big data” platforms like Hadoop, a similarly focused approach to the scope of analysis can be instrumental in reducing the volume of information necessary for the eDiscovery process. A company may employ a big data platform for a variety of reasons. Some of the most common reasons are speed, capacity, and scalability. This can include the need to track thousands of data points, perform analysis and return a response to a query in a very short amount of time. In the context of discovery however, every single data point within the billions of data points stored on these systems may not be directly applicable. Isolating the specific data elements needed to answer the questions related to an investigation often avoids large-scale productions from these platforms. In general, many investigations do not hinge on the individual data points, but on trends present in the data or subsets of data that may be classified as outliers. By taking a focused approach and performing any required analysis on the platform itself, the final output will usually be one that can be managed within a standard database or review platform.

Data review/correlation issues

Once data sources have been inventoried, and manageable subsets of information have been exported from all relevant systems of record, what are the next steps? The most significant task for the eDiscovery practitioner when it comes to these nonstandard data types is transformation of the raw data into meaningful information that can support the case - the process of letting the data tell its story.

Nontraditional data types are becoming more prevalent in modern litigation, and while there have always been industries where these data types were contained in a system of record, today it is universal across all verticals. The analysis of such data may be as simple as isolating text messages between two individuals that are significant to an investigation. This allows a review of the transcript of their conversation without the distraction of hundreds of other messages between different parties. It may also involve comparisons of data sets from back-ups of different time periods to prove that data points related to an industrial test are being manipulated after the fact. Individual data points may be converted to aggregate totals or used to identify general trends. For example, the trend of prescriptions written by a doctor serving as an expert witness can be used to refute the doctor's stated concerns about a drug's safety profile.

While these types of analysis have always been required in eDiscovery, today there is a greater need to correlate information from multiple sources to get a complete picture of the events taking place. Consider a scenario in which financial traders are chatting on a variety of instant messing tools to discuss a questionable trade. They may jump from platform to platform to prevent the generation of a continuous transcript of their conversation. With different usernames on different platforms, the task of piecing together the entire conversation may be difficult. One possible approach would be to export the contents from the various platforms and normalize the usernames such that an individual can be identified regardless of their username on each platform. Once this is accomplished, creating a chronological transcript of their entire conversation, regardless of the messaging platform, becomes a trivial task.

Similar requirements to correlate information from disparate sources may exist when reviewing an audio file of a company representative talking to a customer. The audio file itself may not contain information required to provide full context to the recording. The metadata related to recording may be in the database used by the recording system, and may contain details such as the call time, incoming line, and the representative that answered the call. Details of the records viewed and modified may exist in audit logs and notes entries from the CRM software used by the representative. When all three sources are considered together, the examiner is able to get a much clearer picture of the actual events. When addressing many nontraditional data types in corporate environments today, it is important to consider alternative sources that through review and correlation may provide additional details and clarity to the investigation. As new technologies and review methods continue to emerge, it becomes important to stay on top of the most efficient methods to integrate nontraditional data types from multiple sources. Doing so will provide all the puzzle pieces with less burden on the reviewer.

conclusion

The advancement in connected technologies and software has created an explosion of nontraditional data sources. Applications available to an increasingly mobile workforce have accelerated this trend, and encouraged companies to quickly adopt these technologies. However, due to their complexity, distributed storage systems and varying levels of access and control, these technologies present many challenges to the eDiscovery practitioner. While many of the tools and techniques used to process traditional data may not be practical for these new data types, there are still steps that can be taken to address these challenges. Proactively, companies can reduce their risk by creating and implementing policies governing employee use of these tools, as well as the storage, retention and control over corporate data. Engaging in the creation of a data map and custom workflows for key data sources can prepare a company for a potential investigation or incident.

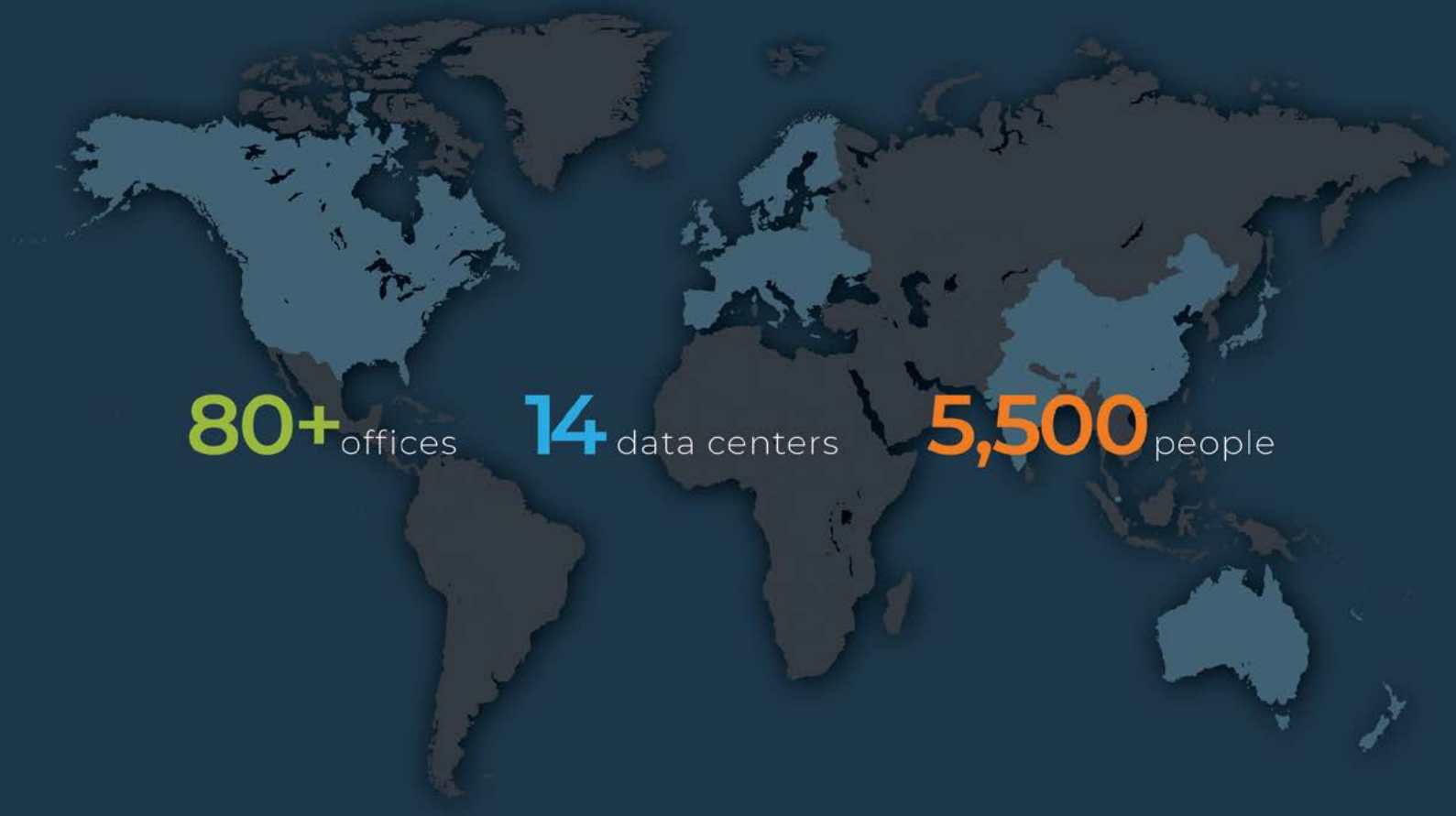
Nontraditional data sources have already become prevalent. Their use, along with the volumes of data they create, will only continue to expand. Preparing for these data types is a critical undertaking for every company, and should be part of any litigation preparedness plan.

About Epiq

Epiq, a global leader in the legal services industry, takes on large-scale, increasingly complex tasks for corporate counsel, law firms, and business professionals with efficiency, clarity, and confidence. Clients rely on Epiq to streamline the administration of business operations, class action and mass tort, court reporting, eDiscovery, regulatory, compliance, restructuring, and bankruptcy matters. Epiq subject-matter experts and technologies create efficiency through expertise and deliver confidence to high-performing clients around the world. Learn more at www.epiqglobal.com.

Class Action & Mass Tort
eDiscovery
Global Business Transformation Solutions
Regulatory & Compliance
Restructuring & Bankruptcy

epiqglobal.com



People. Partnership. Performance.

epiq