Finding a New Safe Harbor

How Technology Can Support Data Protection Compliance

By Michael Becker, Managing Director, Consilio

Consilio®
Together. Stronger.

How Technology Can Support Data Protection Compliance

The nullification of the U.S.-EU Safe Harbor Agreement was a shot heard around the legal world last fall. In the vacuum created by the collapse of this framework, and with new data privacy laws continuously emerging, organizations have been struggling to determine how best to proceed with cross-border transfers for litigation, compliance and regulatory investigations. Data privacy officers, compliance teams and legal counsel are attempting to find a balance between the regulatory framework governing personally identifiable information (PII) and their data transfer obligations. The answer to this challenge may lie in evolving eDiscovery technology.

The Current Data Protection Landscape

In October 2015, the Court of Justice of the European Union issued its decision in *Schrems v. Data Protection Commissioner*, which invalidated the 15-year-old U.S.-EU Safe Harbor Agreement that provided organizations a means—often the only means—for transferring PII across EU borders to the United States. Its replacement, the Privacy Shield, has yet to be adopted by the EU data protection authorities, but it creates stricter obligations, including tighter controls over transfers to third-party data controllers and their agents. In the interim, organizations have limited options for transfers, including adopting model contract clauses that contain EU-approved language, implementing binding corporate rules that require consent from the local data protection authority or obtaining the freely given consent of the data subjects.

The EU is not alone in strengthening protections for personal data. Numerous other nations, particularly in the Asia-Pacific and South America, are bolstering their data protection laws, making it even more difficult for data to cross borders. For instance, although China does not have a single, overarching data protection law, it has enacted a complex series of laws in various sectors that together create a right of privacy. China's Consumer Rights Protection Law governs consumers' personal information, its Decision on Enhancing Internet Information Protection protects any personal data collected on or transferred over the Internet and its sweeping Law of the People's Republic of China on Guarding State Secrets regulates the processing and transferring of sensitive data. Likewise, Russia recently enacted new localization requirements for data controllers who store and process its citizens' personal data, requiring them to first process data in Russia before transferring it out of the country.

Rethinking the Approach to Cross-Border Data Transfers

With so many complex data protection laws in place, and more being adopted all the time, companies with a recurring need to transfer data across international borders must rethink their approach. Two strategies suggested by The Sedona Conference in *Practical In-House Approaches for Cross-Border Discovery & Data Protection* are to "[u]se the Processing stage of discovery as an opportunity to balance compliance with both discovery and Data Protection Laws, thereby

With so many complex data protection laws in place, and more being adopted all the time, companies with a recurring need to transfer data across international borders must rethink their approach.

demonstrating due respect for Data Subjects' privacy rights," and to "consider ways to limit the production of Protected Data" during review. In turn, following this recommendation in tandem with Principle 3 of *The Sedona Conference International Principles on Discovery, Disclosure & Data Protection* can minimize conflicts with data protection laws: the "[p]reservation or discovery of Protected Data should be limited in scope to that which is relevant and necessary to support any party's claim or defense in order to minimize conflicts of law and impact on the Data Subject." In short, one key to success in cross-border eDiscovery is to minimize the data—particularly protected data—for transport.

The Sedona Conference recommends that parties use technology to limit the amount of PII that must cross borders. One suggested approach is to filter using keywords to isolate documents that contain personal data before transferring them elsewhere in the same jurisdiction for review, which could help parties satisfy their obligations to produce information for overseas discovery while still complying with local data protection laws. The Sedona Conference cites an example of searching for terms such as the names of financial institutions to identify potentially sensitive banking information.

However, it is almost impossible to identify all PII in a document set using keywords because it would require the attorneys and teams constructing the search to know in advance what they are looking for, which is rarely, if ever, the case in litigation or regulatory matters. Given the fact that data stores are growing exponentially, and with time of the essence in the heat of litigation and investigations, tools such as duplicate detection, predictive coding, redaction and anonymization must play a role in limiting the need to process and transfer PII.

Duplicate Detection

Because e-mails have both a sender as well as at least one recipient, and collections often span those custodians, exact duplicates are rife across collected data sets. Thus, tools that can identify and eliminate exact duplicates are an essential component because they can greatly reduce the number of documents that organizations must actually review.

Similarly, near duplicate documents, which are documents that contain a high degree of similar content, are estimated to make up as much as one-third of data sets. Technology can lend an assist here as well by allowing privacy review teams to look at nearly duplicate documents together, speeding the privacy review process.

And, select review toolsets can also identify email duplicates – which are email documents that have identical body text. These documents do not deduplicate from exact dupe analysis because their header information makes them unique, even when the text body of the email is identical. When applied for a privacy review, tags can set to automatically propagate to email duplicate records, further reducing the need to review often thousands of email documents.

Predictive Coding

Predictive coding is a type of technology-assisted review that "learns" by examining how senior attorneys code documents and then applies that learning across entire data sets. More specifically, after lawyers code a sample set of documents, the predictive coding algorithm analyzes this training data to discern indicia of for relevance, privilege, "hotness" and the like. The algorithm then classifies every document

in the data set accordingly. Typically, the classification is a ranking of probability that the document belongs to each category: the higher the probability, the more likely the document is relevant, privileged or important.

For cross-border matters that require a privacy review, predictive coding expedites the process by culling nonresponsive documents from a collection and giving counsel a preview of a data set for early case assessment. Experience has shown that predictive coding is mightier than keyword searches when it comes to culling: on average, when parties run predictive coding after keyword filtering, 60 to 70 percent more nonresponsive documents are culled from the population. Moreover, predictive coding allows parties to process more documents quickly, saving precious review dollars and resources. Another key advantage of using predictive coding for privacy reviews is that because predictive coding is driven by statistics, the review process becomes more defensible than when keywords alone are used. This technology gives parties greater confidence that they have trimmed out as much nonresponsive information as possible.

Unfortunately, not all predictive coding software is created equal. Today, most predictive coding software is created by American technologists with U.S.-based trial data. Therefore, the majority of these tools' classifier engines fall short when it comes to evaluating the multilingual data sets typically involved in global matters. For these matters, where privacy is of the utmost importance, counsel should choose a discovery platform with predictive coding technology that supports multilingual document sets without needing to separate the data sets by language and without requiring the additional time and resources to create language-specific computer models.

Redaction

If relevant documents that contain PII must be produced, it may be possible to remove the PII to avoid violating data protection laws before transferring the documents out of the country. Two options are possible: manual and automated redaction.

With traditional redaction, reviewers manually remove all personal information from a TIFF image or a PDF copy of the document to be produced. One difficulty is that this sensitive information often appears in exact duplicate or near duplicate documents or buried within e-mail threads within a data set. This is where review that groups near duplicates (or exact duplicates) together can be an accelerant and where e-mail thread grouping of reviewed sets can also improve redaction speed and consistency. Review administrators must ensure that any redactions are burned into the images so they cannot be removed by opposing legal teams, and they must confirm that metadata and text load files created from the original documents also reflect the redactions so sensitive data is not exposed in metadata fields (such as e-mail subjects or file names).

To address the tediousness and other shortcomings of manual redaction, some review platforms now offer expression-based searches. These tools can automate searches for certain programmable combinations of text, such as e-mail addresses or numeric combinations that represent telephone, employee identification, social security or bank account numbers—which

points review teams to the documents that contain private data. Instead of searching for a specific term, these tools are looking for anything that might represent private data, such as a 16-character numeric sequence that could represent a credit card number or an eight-letter sequence preceded by three letters and a dash that could represent an account number.

Expression-based searches have their own set of problems, however. These searches are notoriously overinclusive, with a huge number of false positive hits for information that is not private. Moreover, there is no guarantee that the software is finding all of the account numbers, given the number of variations possible and because of the possibility of permutations in the expected character string, whether by a typographical error, faulty optical character recognition or document coding.

Furthermore, parties often forget to search beyond documents' body text for PII, and metadata is often riddled with private data. Expression-based searches need to explore all metadata fields, but this is beyond the capability of many document review platforms. And most tools do not permit users to redact metadata: tools must have certain scripts and processes in place to anonymize or redact the presence of PII in the metadata. As a result, most parties must devise a workflow that allows them to flag the offending metadata attribute for that particular document so it is redacted or anonymized on production export.

Until redaction technology evolves further, quality-control processes are of paramount importance. Parties must undergo a series of trial-and-adjustment searches to fine-tune their expression searches. As they proceed, counsel should create a bank of searches that they can apply to every future matter. Given the risk of revealing PII in the body text or metadata of documents, the bottom line is that few alternatives currently exist, aside from eyes-on review, to protect as much PII as possible.

Automated Anonymization

Another method for complying with data protection laws is anonymizing any PII in the data set. With anonymization, personal identifiers are permanently and completely deleted from a document. For example, a producing party could anonymize employee phone numbers into one single business phone number, or data about employee nationalities could be aggregated into showing the number of employees who represent each nationality. Another useful tool is pseudonymization, which still removes all identifying information but retains the links between multiple records pertaining to the same person.

Unfortunately, today's anonymization tools suffer from flaws similar to many redaction tools currently on the market. For instance, if the PII does not match the expected pattern that the anonymization technology is searching for, whether by typographical errors or otherwise, the tool could fail to identify it. Furthermore, privacy advocates may contend that attempts to remove or modify PII simply whitewash legitimate privacy concerns without failing to address them. Finally, opposing counsel are likely to challenge the integrity of any data manipulated for anonymization.

Recommendations for the Future

As this white paper reveals, current eDiscovery technology is not a silver bullet when it comes to addressing the risks of PII; therefore, parties must also create defensible privacy review processes. To be most effective, these processes must begin with a proactive information governance protocol implemented by an established eDiscovery specialist with local roots.

By better managing their PII before litigation or regulatory investigations arise—and when they are not under the duress of eDiscovery—organizations can develop a sound understanding of their data at rest: where their data resides, what kinds of PII it contains and what its risk profile is. The only way for organizations to inventory

their data is through customized, expression-based searches. And to conduct a thorough search, the expertise of an eDiscovery team with a strong local presence will be required.

The key is to find a service provider with experience handling data in the country where it originated so it is aware of the risks and has the optimal technology, infrastructure, bandwidth, workflows and best practices to manage data in that region. The right team will be able to bring its servers, software, project managers and experts to bear for the company or law firm.

Technology alone cannot solve the quandary organizations face when they must handle PII in cross-border litigation. Only by partnering with a seasoned provider with the right knowhow and workflows to exploit the power of technology can organizations create a process that they can feel confident presenting to opposing counsel and data protection authorities.

About the Author

As **Managing Director at Consilio**, Michael Becker supports international financial and industrial companies and law firms with comprehensive planning and management of eDiscovery, computer forensics and managed-reviewed solutions. Michael has extensive expertise in the design, development and implementation of solutions in antitrust, fraud, compliance and corruption procedures, litigation, arbitration and internal investigations with specialized consideration of international data protection regulations and legal process outsourcing. Mr. Becker is a currently licensed attorney in Munich and speaks German, English and Italian.

Consilio®
Together. Stronger.