Professional Perspective

Using Data De-Identification to Protect Companies

Robert D. Keeling, Ray Mangum, and Michele L. Aronson, Sidley Austin LLP

Bloomberg Law

Using Data De-Identification to Protect Companies

Contributed by Robert D. Keeling, Ray Mangum, and Michele L. Aronson, Sidley Austin LLP

Many companies hope to benefit from amassing large amounts of data by mining it for market insights, creating internal business models, and supporting strategic, data-driven decisions. But as companies collect and store increasingly enormous volumes of data, they may unknowingly take on significant legal risks, including potential violations of data privacy laws and increased exposure to U.S. litigation discovery obligations. One way that businesses can mitigate these risks is to de-identify the data they collect and store.

Done properly, data de-identification can minimize risks to privacy interests—e.g., in the event of a data breach— and can help ensure that companies comply with many privacy laws. It can also reduce the likelihood that a company will need to collect, review, and produce such data in discovery as part of litigation or a government investigation.

This remains an emerging area, however, with no bright-line rules. Consequently, it is important for companies to understand how de-identification changes data and what those changes may mean in the context of a party's discovery obligations.

This article offers several guiding principles to assist companies in developing data de-identification practices that will reduce the risks associated with large-scale compilations of data.

Understand Data De-Identification

Data de-identification is the process of identifying and then removing or manipulating identifiers that point to a specific person or entity within a set of data. There is a spectrum of available approaches for transforming identifiable information, as summarized below:

- **Pseudonymous**
  - Direct identifiers have been removed or manipulated
  - Indirect identifiers remain intact
- **De-Identified**
  - Direct identifiers have been removed or manipulated
  - Known indirect identifiers have been removed or manipulated
- **Anonymous**
  - Direct identifiers have been removed or manipulated
  - Known indirect identifiers have been removed or manipulated
  - Additional mathematical/technical guarantees to prevent re-identification

See A Visual Guide to Practical Data De-Identification, Future of Privacy Forum. Many businesses already de-identify the data they collect to comply with various privacy laws, such as the GDPR or HIPAA. In addition to protecting consumer privacy, data de-identification may also protect this kind of data from discovery in U.S. litigation. That's because proper de-identification of the original data set fundamentally alters it, creating a new data set that is distinct from what was originally collected. As the data set is de-identified, the original data is effectively deleted—and a copy of the original data should not be retained—making the original data no longer available.

When data de-identification is done as part of a company's routine business practice, there is a strong argument that the original data would not be discoverable in litigation. De-identification changes the nature of the data such that, like data destruction, the original data no longer exists for purposes of subsequent litigation. Accordingly, companies should typically de-identify data in the ordinary course of business and in compliance with any applicable records retention or preservation requirements. Relevant data should not be de-identified where, for example, litigation is pending or reasonably anticipated.

Make Re-Identification as Difficult as Reasonably Possible

When the purpose of de-identifying data is to mitigate the risks to consumer privacy and to reduce the risk of burdensome discovery obligations, companies should take steps to make re-identification as difficult as reasonably possible. Doing so will provide the most protection for the original data set. For example, if re-identifying the data is burdensome, a company involved in litigation will have a valid argument that producing such data through discovery is not "proportional to the needs of the case" under Federal Rule of Civil Procedure 26(b)(1) and that the data is "not reasonably accessible because of undue burden or cost" under Rule 26(b)(2).

Considering these factors is particularly important to the extent the company envisions using pseudonymization. If the company retains a key that can be used to easily decode the pseudonym, such data is less likely to be secure in the event of a data breach, and it is less likely to be protected from discovery. Indeed, guidance issued by the Office of Human Research Protection (OHRP) in the U.S. Department of Health and Human Services advises that pseudonymization that can be easily reversed if considered only to have been "coded," not anonymized. See U.S. Dep't of Health and Human Servs., Coded Private Information or Specimens Use in Research, Guidance (Oct. 16, 2008).

Beyond Simple Suppression of Identifiers of Entities & Individuals

Especially with respect to structured data—i.e., data in a standardized format, such as databases—de-identification techniques that go beyond simple removal or substitution of individual identifiers may provide an additional layer of protection. These methods also alter the data associated with such identifiers and could include perturbation (adding noise to the original values), data swapping (exchanging values between records), and generalization methods (replacing specific values with more general ones, such as replacing an age with an age range).

De-identifying data in this way can protect individuals' private information, while still allowing companies to capitalize on the benefits of the business intelligence the data provides.

From a litigation standpoint, de-identifying using only simple removal or substitution of entities and individuals can be problematic, as a court may view such data as the functional equivalent of a redacted document—i.e., although the other party cannot see the data, it is actually still there. By contrast, to the extent the data associated with such entities and individuals has been actually modified or otherwise manipulated by the company, there is a strong argument that the original data no longer exists and that the de-identified data is, in fact, an entirely different set of data.

De-Identify Data at the Entity & Individual Level

Companies often collect datasets from multiple business clients, that in turn collect personally identifying information from their individual customers. For example, a company may provide auditing services to businesses that have collected data from their own business customers. In these scenarios, it may be helpful to de-identify not just the personally identifying information linked to the individual consumers' data—i.e., the information within the datasets—but also any identifiers that could point to the entity that originally collected that individual data.

To accomplish this, a company can store the data in a manner that makes it difficult or impossible to later associate the de-identified dataset with the business that provided it. For example, it may be advantageous to commingle the de-identified data of multiple business clients. A company may also want to avoid maintaining records that would make it easy to associate the data with a particular business client by relying on "external" data sources—e.g., a log file showing when data was transferred from a particular client FTP or when it was loaded into a data lake.

Taking these steps would not compromise the utility of the data for business intelligence purposes but would protect the privacy interests of business clients and their business consumers, and would make the information less discoverable in litigation.

Meet or Exceed the HIPAA Safe Harbor

Companies with health-care data should carefully consider the identifiers that should be de-identified in the data sets they collect and store. The "Safe Harbor" for complying with the HIPAA Privacy Rule, 45 C.F.R. § 164.514(b)(2), provides helpful guidance, setting forth 18 specific identifiers that should be removed in order for the data to be deemed de-identified:

- Name
- Address (all geographic subdivisions smaller than state, including street address, city, county, and zip code)
- All elements (except years) of dates related to an individual (including birthdate, admission date, discharge date, date of death, and exact age if over 89)
- Telephone number
- Fax number
- Email address
- Social Security number
- Medical record number
- Health plan beneficiary number
- Account number
- Certificate or license number
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web URL
- IP address
- Finger or voice print
- Photographic images
- Any other characteristic that could uniquely identify the individual

This rule addresses de-identification from a privacy perspective and is focused on health-related information, but the list of identifiers—particularly direct identifiers such as name and Social Security number—is a useful guideline for discovery protection, too. Although the 18 identifiers in the HIPAA Safe Harbor are not an exhaustive list, removing references to these identifiers may sufficiently alter the data to make the burdens of producing it disproportionate to the needs of the case under Federal Rule of Civil Procedure 26(b)(1), or even to render the data not reasonably accessible under Rule 26(b)(2). Thus, adopting a routine business practice of data de-identification that removes these identifiers is a significant step towards limiting discovery obligations.

Contemporaneously Delete Original Source Data

Efforts to de-identify data may only go so far if the original source data is separately retained. If a company can somehow retrieve the original source data or can recreate it with relative ease, courts may order the production of re-identified data, concluding that the burdens of production would not be disproportionate to the needs of the case and, further, would not render the data inaccessible.

Accordingly, companies should strongly consider developing standard policies and procedures to ensure that original source data is contemporaneously deleted at the time of de-identification or promptly thereafter. Of course, before doing so companies must ensure that they are complying with any independent legal obligations regarding data preservation, including any legal hold requirements for reasonably anticipated or active litigation and all applicable statutory and regulatory requirements, such as the Sarbanes Oxley Act § 802, 18 U.S.C. § 1519.

Apply Sound Information Governance Practices

Even though de-identification affords significant data privacy protections and reduces the risk that data will be discoverable in litigation, even de-identified data–like ordinary data–should be retained only for its useful lifespan. For example, if a company determines that de-identified data from only the most recent five years is relevant for business purposes, it should consider implementing an automatic "roll-off" of older data to ensure that unneeded data is regularly purged without human intervention. Sound information governance practices should continue to apply even to de-identified data.