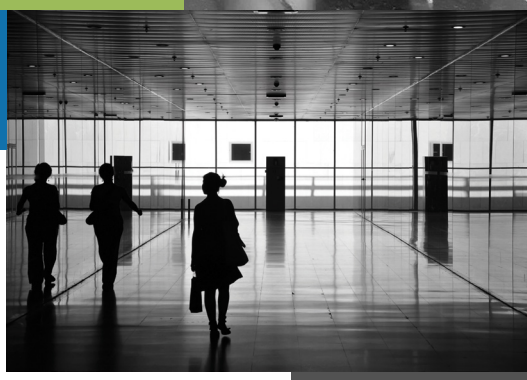
PRIVACY ANALYTICS

WHITE PAPER

De-identification of Unstructured Data

According to one organization, 80 percent of medical record data will be unstructured within two years.[1] Unstructured data can be a critical source of new insights, innovation and knowledge for research hospitals and organizations, medical device companies, insurance companies and medical claims processors, among others. Here's how to unlock it.



PRIVACY ANALYTICS

a QuintilesIMS company

The Proliferation of Unstructured Data

EXAMPLE OF LEXICON IN USE

The Biomedical Translational Research Information System (BTRIS) at the National Institute of Health (NIH) Clinical Center, a biomedical research facility and an agency of the United States Department of Health and Human Services (DHHS), plans to de-identify unstructured text data from more than 400,000 patients for research purposes using Privacy Analytics' Lexicon software. They intend to augment the data currently available in de-identified format within its BTRIS data repository. The addition of unstructured text data without personal identifiers to the repository will allow researchers access to NIH Clinical Center clinical documentation from 1976 to the present. Access to clinical documentation in addition to structured data in de-identified form allows researchers to test hypotheses for new research, confirm potential sample sizes for proposed research and find collaborators for cross-disciplinary research studies.

Lexicon de-identifies personal information, such as direct and indirect identifiers found in physician notations in structured databases and medical devices, residing in text and XML formats. Statisticians and data analysts can now de-identify hundreds to millions of records concurrently, while at the same time gaining the benefit of analyzing this information in compliance with HIPAA and other legal requirements.

Social media has long been associated with the astonishing growth of unstructured data. It has gotten the lion's share of attention by media and industry, which highlight the exponential growth in volume and virtues of insights to be gained. Gartner Research Inc., a global industry research firm, reports that social media revenue is rapidly growing on a global basis, with a projected Compound Annual Growth Rate (CAGR) of 23 percent over five years, from an estimated (US) $11.8 billion in 2011 to $33.5 billion by 2016.[2]

And yet, while the mining of social media, performing sentiment analysis may indicate an individual's desire to advocate and even purchase a brand, unstructured data's wider societal benefits may well reside in the unassuming text fields of electronic health records, medical devices, and discreet online health forums.

Consider Electronic Health Records or EHRs. As just one source of unstructured data, they represent a rich repository of free text. The free text can exist as fields in a database, standardized XML files that are exported to allow data exchange, or as simple text file dumps from medical records or medical devices. The value of this data, therefore, resides in the richness of its analytic depth, of patient narratives, clinical summaries and transcriptions that highlight the rationale for healthcare decisions and ultimately their costs.

Mounting healthcare costs are well documented and beyond our scope here. They are, however, driving a need to marshal greater efficiencies within healthcare organizations and clinical research, decisions ultimately justified by more textured layers and understanding of patient level data. Moreover, many healthcare organizations, such as insurance claims processors, have multiple legacy systems, making it difficult to analyze the totality of an individual's structured data, let alone unstructured.

DE-IDENTIFICATION OF UNSTRUCTURED DATA

Secondary data use, the sharing of personal information outside of direct healthcare delivery, throws up an even more significant challenge: protecting personal healthcare information in compliance with HIPAA and other jurisdictions' legal requirements. Many jurisdictions now also have data breach notification laws. The costs of breach notification are high, representing on average $200 per individual and as much as 7 million per organization.[3] Yet, the argument for secondary data use in healthcare, both of structured and unstructured, is too compelling, despite its risks. These risks can be readily mitigated with proper de-identification.

Secondary data enriches healthcare research and quality of care and delivery, while accelerating the marketing of medical innovations. It can also improve the relative performance of providers, helping to lower overall healthcare costs. Further leveraging data for secondary purposes creates revenue streams for the collection and sale of healthcare data to third-parties, such as drug and device manufacturers, governments, payers and researchers.

The challenge for many organizations leveraging unstructured data for secondary use, however, is systematically evaluating the relative risk of sharing data while ensuring its de-identification allows for high quality data analysis. In short, how should organizations establish standard methods that enable a repeatable, scalable and compliant analytic pipeline that can leverage unstructured data for secondary use? Our view is that a systematic approach is required, one which automates de-identification and risk analysis, and which is governed by rigorous compliance practices. The underlying management of structured and unstructured health data must incorporate de-identification as a necessary best practice when that data is used and disclosed for secondary purposes. As a result, organizations can establish a common approach and rigor for enabling the secondary use of data, critical pillars that:

- Allow the configuration of anonymization for patient level data analysis without compromising privacy and risking costly breaches;
- Ensure de-identified data has analytic usefulness; and,
- Enable analysis of the total patient health experience, to compile a complete picture of this experience from multiple data sources and types.

Secondary data enriches healthcare research and quality of care

Considerations

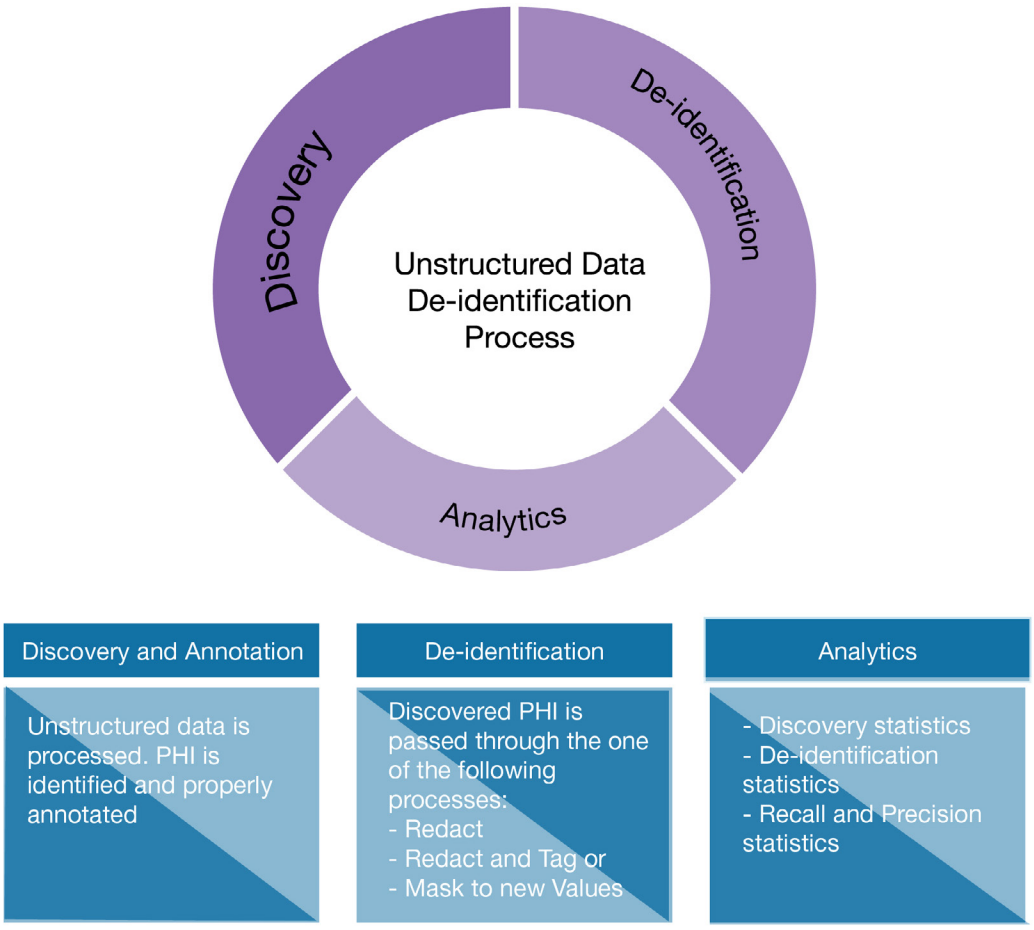


Figure 1. Unstructured De-identification process

There are five considerations to successfully leveraging unstructured data for secondary use.

Consideration 1: Assessing Your Organization's Readiness for Secondary Use

Privacy Analytics has created a maturity model framework that gauges the level of an organization's readiness and experience with respect to anonymization in terms of people, processes, technologies and consistent measurement practices.

The De-identification Maturity Model (DMM) is used as a measurement tool and enables an enterprise to implement a fact-based improvement strategy.[4] DMM is intended to serve a number of purposes: (1) it can be used by organizations as a yardstick to evaluate their de-identification practices; (2) it provides a roadmap

for improvement, helping organizations to determine what they need to do next in order to improve their de-identification practices; and (3) it allows different units or departments within a larger organization to compare their de-identification practices in a concise and objective way.

Organizations that have a higher DMM maturity score are considered to have better and more sophisticated de-identification practices. Higher maturity scores indicate that the organization is able to: (1) defensibly ensure that the risk of re-identification is "very small"; (2) meet regulatory and legal requirements; (3) share more data for secondary purposes using fewer resources; (4) share higher quality data that meets the analytical needs of users; (5) de-identify data through consistent practices; and (6) better estimate the resources and time required to de-identify a dataset.

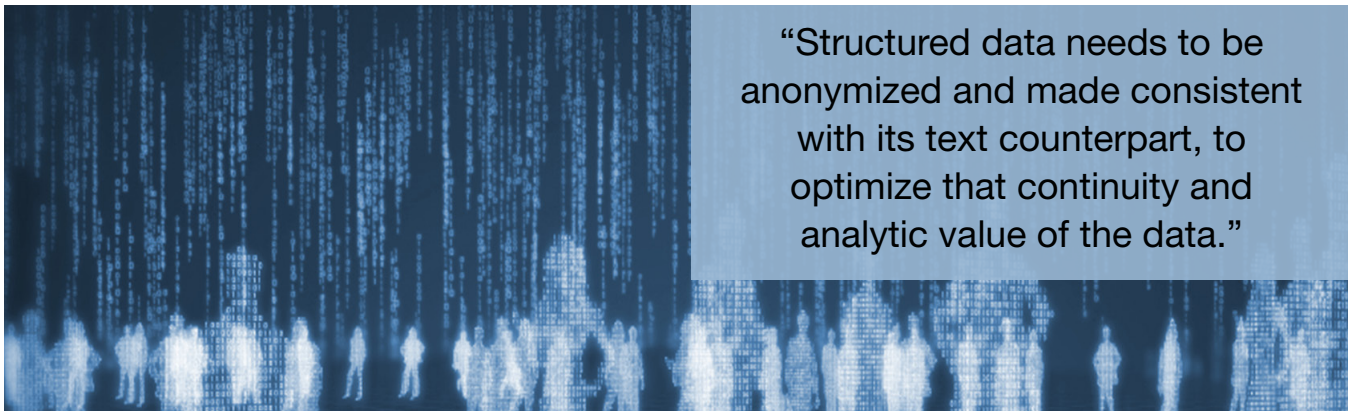
In the context of unstructured data, the DDM allows data custodians to take a holistic approach to all their data types and manage the re-identification risks enterprise-wide. There is no need to have one approach for structured data and another approach for unstructured data.

Consideration 2: Anonymizing Unstructured Data

Our approach to the de-identification of unstructured data has three components: (1) information extraction or discovery; (2) de-identification; and (3) analytics. In the discovery phase, unstructured data is passed through a natural language processing (NLP) engine. This step finds personal identifying information elements in the original document and tags (or annotates) them.

These annotations are then passed through the de-identification phase where the user has an option to either:

- Redact: Replaces all identifiers with a special set of characters such as "***"
- Redact and Tag: Replaces each identifier with an indexed tag, such as [Firstname, 1] or [City, 5]. The index allows you to match the same name across the same document, or even across multiple documents.
- Randomize and Replace: Uses a dictionary



or gazetteer list of names to replace each identifier with a random text or values from the list according to the tag type. For example, a Firstname tag value will be replaced with a name from the gazetteer list of first names.

Consideration 3: Enabling Data Consistency

To ensure data consistency, software can be used on internal indexing databases to keep track of the tag indexes and maintain referential integrity of the data.

For example, in a typical relational data base with many records, a patient may have multiple visits to a hospital and have different discharge records and notes. Referential integrity ensures that the same identifier, for instance a first name like Bob, is replaced with the same name (e.g., John) across all files and text fields within an individual's input folder.

Similarly, structured data needs to be de-identification and made consistent with its text counterpart, to optimize the continuity and analytic value of the data. Any modifications to the structured and unstructured data should remain consistent across the entire dataset. As a result, data analysts can match masked data to corresponding de-identification unstructured text, to ensure the analyses of identical values.

Consideration 4: Ensuring Compliance to Mitigate Risk

Compliance standards are critical to establishing that a de-identified dataset has a "very small" risk of re-identification. When determining re-identification risk in a structured and unstructured dataset, data analysts must consider recall and precision. Recall is essentially how many personal identifiers are detected, while precision measures to what extent text (words or phrases) that are not personal identifiers are actually tagged by software as personal identifiers: in other words, the higher the precision, the less distortion the data has for analysis.

For example, certain abbreviations in clinical notes look very similar to postal or zip codes. In some cases, a less than robust de-identification solution could treat the abbreviations as unique identifiers and redact them, which would limit the analytic richness of the data and may distort the clinical information.

Our approach to recall is an all or nothing approach. If one has a document that finds the name "Bob" 90 percent of the time, then the document is still identifiable. This means that recall is zero not 90 percent, because 10 percent of the instances of "Bob" in the document are not found.

The threshold, therefore, must be extraordinarily high, well beyond accepted academic and industry practices. In some instances, instead of having a fixed 90 percent threshold value, the threshold used for recall may depend on the results of a risk analysis. This risk analysis would follow Privacy Analytics' methodology, which is used for de-identifying structured data as well.

Consideration 5: Establishing a Scalable and Repeatable Anonymization Processes

When we think about scaling and repeatable processes, we often think of technology. Indeed,

software plays a key role in automating de-identification. Privacy Analytics' Lexicon software can handle very large databases of 100's of millions of records.

For unstructured data, Lexicon consumes millions of documents in a real-time streaming environment, de-identifies them, and places the exported data in a destination location. Additionally, it can pull data from multiple data sources through its API, automatically integrating with customers' IT environments – regardless of their complexity. For statisticians and data analysts, then, de-identification software allows them to automate complex analytical tasks, such as discovery, de-identification and masking of personal information. Such software further mitigates the risk of re-identification by detecting exposure of personal information, while also determining its relative analytic quality. And lastly, it optimizes the value of data assets by maintaining the relationship of masked and de-identified values for more granular, higher quality analyses.

But that's only half of it. Technology for technology's sake does not drive an organization forward. It's an enabler of improved productivity, of higher quality analysis. With de-identification software, such as Lexicon, privacy and compliance officers strengthen the protection of their organizations' data assets by extending de-identification to structured and unstructured data. This allows organizations to create an enterprise-wide basis for repeatable, scalable process. It also ensures that secondary data use is compliant with HIPAA and other legal requirements and that internal organizational policies and procedures covering privacy and personal information are aligned and managed consistently throughout their organizations.

Many organizations have been collecting unstructured data for some time. To date, however, gaining analytic utility from unstructured data has proven challenging, as most organizations have inadequately address its use and disclosure for secondary purposes.

As we have noted, the potential benefits from being able to do more with unstructured data are significant, from research, public health, to commercial and policy making applications. Its growth – in all forms and uses – is growing

CONTACT US

251 Laurier Ave W
Suite 200
Ottawa, Ontario, Canada
K1P 5J6

Phone: 613.369.4313

www.privacy-analytics.com

sales@privacy-analytics.com

exponentially and represents significant opportunities. Key to these opportunities is an automated, scalable practice of de-identification. Software is vital, but so too is underlying assessments and process associated with its management – practices that are managed using scalable, repeatable processes.

Privacy Analytics has developed solutions, including software and professional services, to de-identify unstructured data so that it can be used and disclosed. Our de-identification solutions enable the secondary use of structured and unstructured data together, so that all of a data custodian's information assets can be leveraged in a consistent way. In combination with the Privacy Analytics' risk management methodology and maturity model, a roadmap for de-identifying data can be developed taking into account the organization's technical, process and resource capacities.

To learn more about our Lexicon software, download the datasheet.

Sources:

1. Kalakota, Ravi, Reference to Gartner Research Inc., http://practicalanalytics.wordpress.com/2013/10/23/market-sizing- analytics-and-big-data/

2. ZDNet, April 9, 2013, "Within Two Years, 80% of Medical Data Will Be Unstructured," http://www.zdnet.com/article/within-two-years-80-of-all-medical-data-will-be-unstructured/.

3. El Emam, K., Guide to the De-identification of Personal Health Information. CRC Press (Auerbach), 2013

4. El Emam, K., Hassan W., The De-identification Maturity Model.