edios
Innovative Software Solutions

eInfoGuard®

Data De-Identification Solution

Data Masking

White Paper

eInfoGuard Data Masker Solution Ver 1.0

Abstract & Implementation

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule mandates the de-identification of specific types of Protected Health Information (PHI) for covered entities and their business associates.

This document discusses each of the items listed in §164.514(b) of the Privacy Rule's Safe Harbor de-identification standard and illustrates various data masking / anonymization techniques which can assist in satisfying that requirement.

Edios eInfoGuard is a comprehensive solution for Data De-Identification of Development, Test, Training or Staging Databases by either masking or anonymizing the field values. **Edios eInfoGuard** supports multiple databases – Oracle, MS SQL, Sybase and MySQL.

In **Edios eInfoGuard**, a Profile can be created for Data De-Identification of any database. A Profile can have single or multiple rules and in one rule, single or multiple field values can be de-identified by using various techniques. Once the Profile is created for any database, it can be saved for future use for Data De-Identification of the same database when Production Database is restored in Development, Test, Training or Staging Environments.

Edios eInfoGuard Data De-Identification solution's various Data Masking & Anonymization Techniques are explained below by taking different Techniques, Field Names & Data Type examples.

This document is specifically focused on the practical issues and techniques involved in the preparation of de-identified HIPAA compliant databases. Specific worked examples of data masking practices which can assist with meeting the mandated Safe Harbor data de-identification requirements are provided.

Edios eInfoGuard Data Masking & Anonymization Comprehensive Solution is used to illustrate the example of various masking techniques in this document. The worked examples in this document are loosely based on the HIPAA data de-identification guidance provided on the Office for Civil Rights (OCR) website located at

http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

Compliance to HIPAA

HIPAA Privacy Rule

The HIPAA Privacy Rule is intended to protect information (however held) which may identify an individual. In general the requirement for inclusion as HIPAA sensitive information is any data, including a demographic detail, which relates to:

- the individual's past, present, or future physical or mental health or condition,
- the provision of health care to the individual, or
- the past, present, or future payment for the provision of health care to the individual, and that identifies the individual or for which there is a reasonable basis to believe can be used to identify the individual. Protected health information includes many common identifiers (e.g., name, address, birth date, Social Security Number) when they can be associated with the health information listed above.

The determination of which types of data are considered to be PHI data essentially reduces to a determination of whether the data contains any information which may identify an individual and also contains the associated health data content.

HIPAA Privacy Rule Data De-Identification Methods

There are two implementation specifications a HIPAA covered entity can follow to meet the Privacy Rule data de-identification standards. These two specifications (called methods) are covered by sections 164.514 parts (b) and (c) of the Privacy Rule and are simply named the Safe Harbor Method and the Expert Determination Method.

The Expert Determination Method basically reduces to a "judgment call" by a suitably knowledgeable and qualified person who, in their professional opinion, believes the operations performed on the data have rendered it into such a state that there is a very small risk that anybody viewing the data could use it to identify an individual. The Safe Harbor Method prescribes a list of identifiers related to an individual (or relatives, employers, or household members of the individual) which should have data de-identification operations performed on them.

Data De-Identification Meaning

Data De-Identification is the replacement of existing sensitive information in Development, Test, Training or Staging Databases with information that looks real but is of no use to anyone who might wish to misuse it. In general, the users of the Development, Test, Training or Staging Databases do not need to see the actual information as long as what they are looking at looks real and is consistent.

It is important to be aware that Data De-Identification is appropriate to more than just personal details – sometimes business confidential information is appropriate for masking as well. For example, it may be desirable to prevent quarterly sales figures for some products being present in an outsourced Test Database.

Data De-Identification is not the same thing as restricting the visibility of information in Production Databases from people who are not authorized to see it. In that situation, the data is actually present in the database and is simply not visible to the unauthorized. There are many good and justifiable reasons for taking this approach in a Production System, but adopting a “data is present but hidden” approach to the protection of data in Development, Test, Training or Staging Databases is a recipe for trouble. The reason is that strict controls are in place in Production Databases and these can present a carefully managed view. Developments, Test, Training or Staging Systems are different. Typically, they are an environment in which access is usually much wider. Information is visible to more people and those people often have greater privileges and low level access. From a data visibility standpoint, a Development, Test, Training or Staging System in which the data is present but hidden is a system which sooner or later will expose its data.

In general, a reasonable security assumption is that the more people who have access to the information, the greater the inherent risk of the data being compromised. The modification of the existing data in such a way as to remove all identifiable distinguishing characteristics yet still usable as a Development, Test, Training or Staging system can provide a valuable layer of security. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule mandates the de-identification of specific types of Protected Health Information (PHI) for covered entities and their business associates.

Why Data De-Identify

Legal Requirements

The regulatory environment surrounding the duties and obligations of a data holder to protect the information they maintain are becoming increasingly rigorous in just about every legal jurisdiction. It is a pretty safe assumption that the standards for the security and maintenance of data will become increasingly strict in the future.

Loss of Confidence and Public Relations Disasters

It can reasonably be said that if a data escape happens at any organization, then the formal legal sanctions applied by governmental bodies is not the only problem that will be faced. Possibly it may not even be the biggest of the immediate worries. Inappropriate data exposure, whether accidental or malicious, can have devastating consequences. Often the costs of such an event, both actual and un-quantifiable can far exceed any fines levied for the violation of the rules. For example, what will it cost the organization if potential customers are not willing to provide sensitive information to the company because they read an article about a data escape in the newspaper. Dealing with the public relations aftermath of seeing the companies name in the press will not be cheap. It also does not take much imagination to realize that senior management are not going to be happy about having to give a press conference to reassure the public. The public relations costs of a data escape usually far exceed the sanctions levied by governmental organizations.

Malicious Exposure

Most people think the major risk to the information they hold is external entities (and organized syndicates) out to break in and steal the data. The assumption then follows that protecting the network and firewalls is the appropriate and sufficient response. There is no denying that such protection is necessary – however it has been shown that in many cases the data is stolen by malicious insiders who have been granted access to the data. No firewall can keep an insider from acquiring data under such circumstances. However, by reducing the number of databases with unmasked information, the overall risk of exposure is mitigated. The external hackers, if they get through the network security, will have far fewer useable targets and a far greater proportion of the inside personnel will have no access to the real data.

Accidental Exposure

The risk of accidental exposure of information is often neglected when considering the security risks associated with real test data. Often it is thought that “there is no point in masking the test data because everybody has access to production anyways”. Not so, the risks associated with an accidental exposure of the data remain. Often just masking the most sensitive information (credit card numbers, customer email addresses etc.) is enough to somewhat mitigate the damage associated with accidental exposure and the masked databases remain just as functional.

Common Fields & Identifiers for De-Identification

The following are few examples of common Fields &Identifiers that can be de-identified using Edios eInfoGuard:

- Names
- Geographic Locations
- Date Fields directly related to an individual - Birth Date, Admission Date, Discharge Date, Death Date etc.
- Postal Addresses
- Telephone Numbers
- Fax Numbers
- Email Addresses
- Social Security Numbers
- Medical Record Numbers
- Health Plan Beneficiary Numbers
- Account Numbers
- Credit Card Numbers
- Certificate/License Numbers
- Vehicle Identifiers and Serial Numbers - including License Plate Numbers
- Device Identifiers and Serial Numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) Address Numbers
- Healthcare Records
- Any other unique identifying number, characteristic, or code which identifies an individual

Data De-Identification Techniques

Substitution or Replacement is a Data De-Identification Technique to substitute or replace any field value based on some business logic. This technique can be applied to various fields – First Name, Last Name, Phone Numbers, Credit Card Numbers, SSN, Geographic Locations etc. Substitution or Replacement is very effective in terms of preserving the look and feel of the existing data.

Random Value Replacement (Same Field): A rule can be created in Edios eInfoGuard to replace the field values of any field randomly of any data type (String, Number, Date etc.). For example, Last Name field values can be replaced randomly using this rule:

Before Data De-Identification		
ID	**First Name**	**Last Name**
1	Adam	Pape
2	Alan	Graham
3	Brian	Stagner
4	James	Smith
5	John	Frank

After Data De-Identification		
ID	**First Name**	**Last Name**
1	Adam	Stagner
2	Alan	Smith
3	Brian	Frank
4	James	Graham
5	John	Pape

Random Value Replacement (Different Table &Field): A rule can be created in Edios eInfoGuard to replace the field values of any field of any data type (String, Number, Date etc.) randomly from different Table and Field of same data type. In this case, field values will be replaced randomly from other selected table and field.

Random Value Replacement (CSV File): A rule can be created in Edios eInfoGuard to replace the field values of any field of any data type (String, Number, Date etc.) randomly from field values uploaded from CSV File. In this case, fields value can upload through CSV file and can be used for random replacement.

Static Value Replacement: A rule can be created in Edios eInfoGuard to replace the field values with any Static Value. For example, Person Email Addresses can be replaced with some dummy or test email addresses.

Append Static Value: A rule can be created in Edios eInfoGuard to append any Static Value to any field value as either Prefix or Suffix. For example, Medical Record Numbers can be prefixed with “Test” to indicate the test data.

Masking or Replacement of Particular Character(s): A rule can be created in Edios eInfoGuard to mask or replace any particular character(s). For example, Character “1” can be replaced with Character “X” and Character “2” can be replaced with Character “Y” etc. in Social Security Number.

Masking or Replacement of Particular Position(s): A rule can be created in Edios eInfoGuard to mask or replace any particular position (s). For example, Position 5th Character can be replaced with Character “X” and Position 6th Character can be replaced with Character “Y” etc. in Social Security Number.

Data De-Identification Techniques

Masking or Replacement of Vowel(s): A rule can be created in Edios eInfoGuard to mask or replace all Vowels in a string with any particular character.

Random Shuffling: A rule can be created in Edios eInfoGuard to randomly shuffle any field value. For example, Phone No. "(510) 405-0334" after shuffling may become "(103) 554-3400". In Shuffling, it can also be defined that Special Characters ("Bracket, Hyphen, Comma etc.) are not shuffled and remain at same positions. For example, in above case Brackets & Hyphen remained at same position even after shuffling.

Range Wise Randomization (Number Value): A rule can be created in Edios eInfoGuard to randomize the any Number Value with in defined range. For example, Person Age can be randomized to any random value by defining the range of value between 10 to 60 Years etc.

Range Wise Randomization (Date Value): A rule can be created in Edios eInfoGuard to randomize the any Date Value as a whole value or individual element (Day, Month or Year) with in defined range. For example, Person Birth Date as a whole value can be randomized to any random value by defining the range of value between 01-Jan-1980 to 31-Dec-2015. The individual element "Day" can be randomized by defining the range value from 1 to 28, "Month" can be randomized by defining the range value from 1 to 12 and "Year" can be randomized by defining the range value from 1980 to 2015 etc.

Conditional Data De-Identification: All Rules can be applied to all data set or filtered data set of a table. Edios eInfoGuard provides the User Interface to define any filters on the data by applying various conditions.

Edios eInfoGuard also provides the functionality to replace the field value with same or different values as mentioned below:

Single Row Replacement: Using this option, same field value will be replaced with different value for all records. For example, there are 5 records with City "Chicago". By using this option, "Chicago" City will be replaced with different value (Dallas, Newark, San Francisco etc.) for all 5 records.

Multiple Rows Replacement: Using this option, same field value will be replaced with same value for all records. For example, there are 5 records with City "Chicago". By using this option, "Chicago" City will be replaced with same value (Dallas or Newark or San Francisco etc.) for all 5 records.

Supports Multiple Databases: Edios eInfoGuard Data De-Identification solution supports multiple databases - MS SQL,Oracle, Sybase and MySQL. The data fields of all data type of these databases can be de-identified using Edios eInfoGuard.

About EDIOS

EDIOS is a forward minded software and technology services company with various Software Products in Healthcare Domain. Our services and solutions help clients implement business strategies, improve performance, increase efficiency, and deliver changes effectively. This includes managing projects, taking responsibility for business processes, and facilitating business growth through innovation and application of technology. Headquartered in Salt Lake City (U.S.), EDIOS combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. Please visit our Web Site www.edios.global

Our Products, Solutions & Services

- ✓ Healthcare Intra-operability Solution
- ✓ EMR, LIMS & Other 3rd Party Systems Integration
- ✓ Registry Management System (RMS)
- ✓ Electronic Patient Referral System (EPRS)
- ✓ Mobile App Development (Android, iOS& Windows)
- ✓ Web App Development (Java,ASP.Net & PHP)
- ✓ OS (Windows & Linux)
- ✓ Database (MySQL, MS SQL, Oracle & Sybase)
- ✓ Customer Support (Level 1, 2 & 3)
- ✓ Data Analytics

edios
Innovative Software Solutions

North America
4683 Garden Spring Ln
Salt Lake City, UT 84117,
U.S.A.

Europe
Rue du Littoral 23
CH-2025 Chez-le-Bart
Switzerland

Asia
C-139, Ind.Area, Phase -VIII
S.A.S. Nagar,Mohali-160059,
Punjab,India