IRI
Total Data Management

The Case for
DATA MASKING
& IRI Capabilities White Paper

Updated January 2023

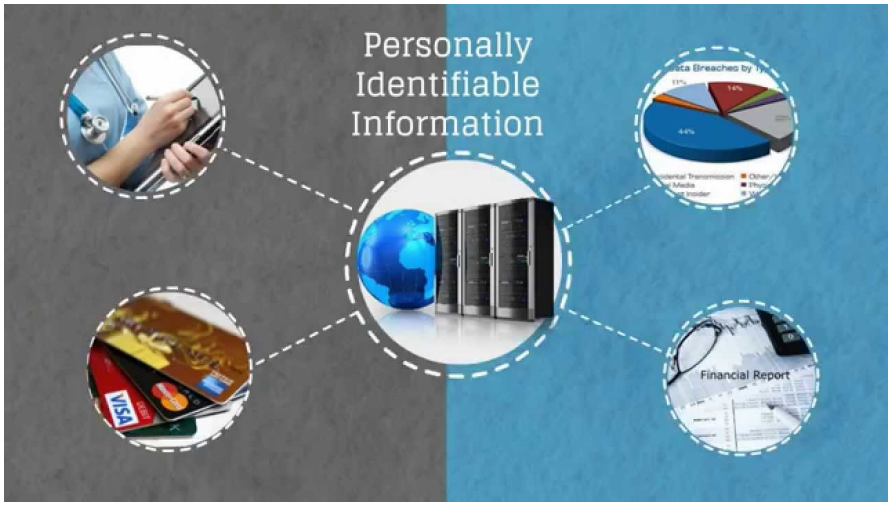
Table of Contents

Introduction

Amid non-stop reports of data breaches and a growing regulatory environment for personally identifiable information (PII) worldwide, multiple technology solutions and compliance services have arisen to address PII protection.

Logical security through encryption in one form or another is a common approach denominator, but most commercial encryption applications are limited by platform, algorithm, ciphertext appearance, implementation complexity, runtime performance, and cost. Moreover, wholesale encryption of data sources and devices removes access to non-sensitive data, renders targets unsuitable for DevOps and testing, and leaves all the data vulnerable to a single decryption key breach.

There is a need for targeted, versatile, and efficient methods to identify, protect, and audit the remediation efforts applied to PII in different sources. Data masking technologies lead the way here.



IRI is a data masking industry pioneer, and more recently expanded the industry's scope to include ancillary functions under the company's own term: Startpoint Security. This term, later defined in more detail, includes the discovery and de-identification of PII and other sensitive data, but also tightly integrated functions and processes associated with them, including data profiling and classification, re-ID risk scoring, test data provisioning, role-based access controls, and audit.

This white paper makes the case for data masking in particular however, starting with the regulatory landscape behind it and the technical aspects of different masking functions which can accommodate different needs (like determinism for uniqueness and reversibility, realism without reversibility, etc.). It then moves into general recommendations before reviewing selected masking functions that help nullify data breaches, comply with data privacy laws, and satisfy many test data use cases.

The paper concludes with a list of applicable IRI data masking technologies reviewed by Gartner, and links to more information. A supplement to this white paper describing ways in which IRI software helps achieve specific COBIT objectives in a larger risk and controls framework is also available on request.

The Compliance Landscape

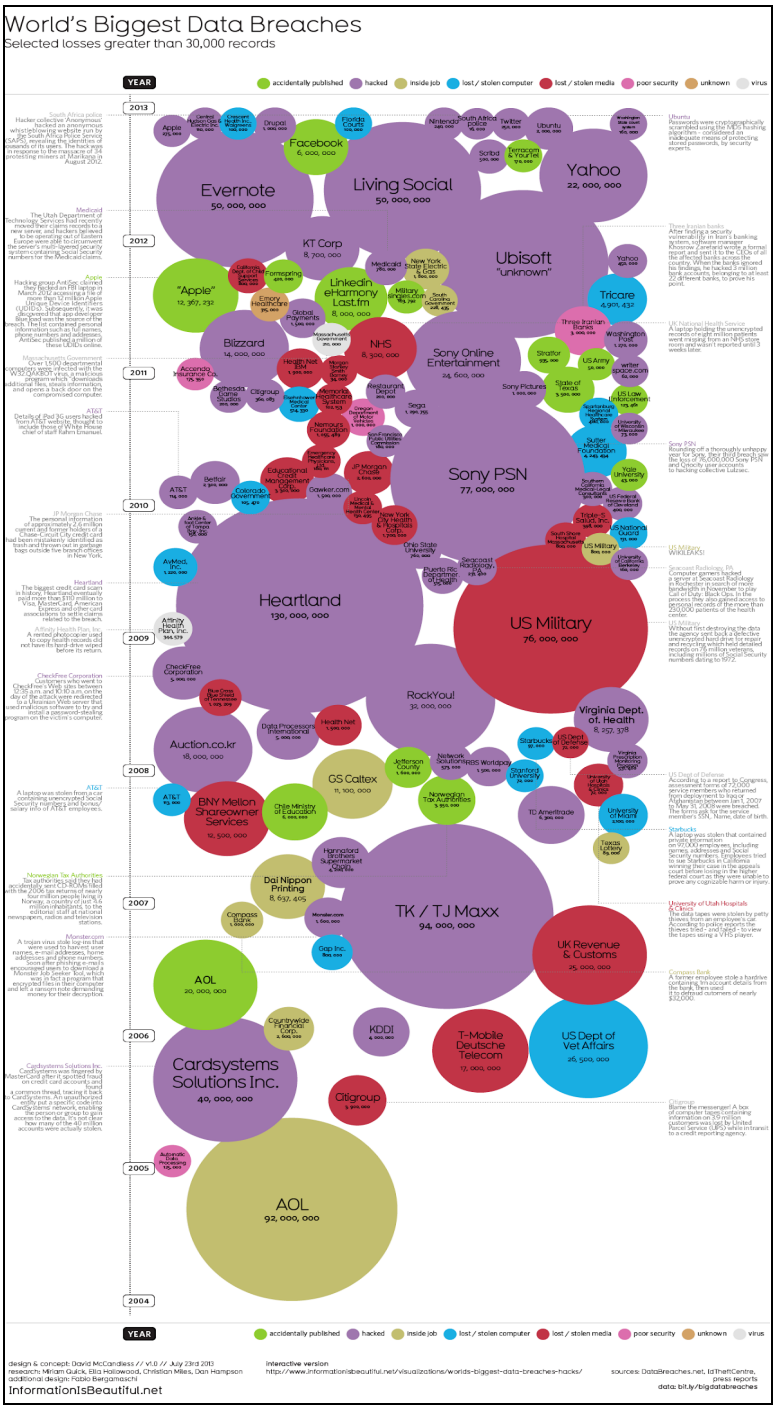
Corporate leaders and information technology executives understand the rewarding possibilities of business intelligence gleaned from their stores of transaction data as well as public sources. Unfortunately, there is a converse, penalty-enforced requirement to protect data from unintended users and uses. Put another way, revenues must also be safeguarded through *safer data*.

A major emphasis is therefore being placed on compliance with government regulations, industry standards, and corporate policies designed to identify and protect data at risk of misuse through disclosure, outsourcing, hacking, theft, and otherwise improper handling.

From an auditing and risk control perspective, executives are scrambling to understand compliance and fund the implementation of controls that will solve problems with data before they surface. From a data perspective, business and IT stakeholders are also discovering that data quality is questionable, and that they often cannot trace data from point A to B without disconnects; i.e., where data starts in the system and ends in reports. Understanding and verifying these data flows are keys to complying with regulations like Sarbanes-Oxley (SOX).

In this compliance era, the common language is now one of risk management. IT departments are now learning that language in many of the same ways business people have always had to – including the hard way. This paper lays out that landscape, and suggests ways to help achieve and verify compliance, while still producing rapid, actionable business intelligence.

Let's begin with some of the key considerations and recommendations surrounding risk mitigation in the post-SOX environment:



1) Data = Risk

Data are always at risk – at risk of hacking, theft, incorrect transmission or modification, being pulled from the wrong sets, and so on. If sensitive data are disclosed, the Privacy Rights Clearinghouse (PRC) posts the incident for the world to forever know what companies allowed its customers' personal information to be compromised. In addition to the embarrassing secrets revealed, lawsuit damages, government fines, and/or data remediation costs can cripple the entire company.

According to the PRC and its Chronology of Data Breaches, roughly one billion personal records have been reported compromised in the United States alone since 2005. Much of the PII obtained by identity thieves include Social Security, driver's license, and various account numbers. The PRC's enormous and still-growing list is also comprised of breaches that do NOT expose such sensitive information in order to emphasize the wide array and prevalence of data breaches.

2) Risk Must Be Managed

Managers must understand the risks data represent and then formulate, execute, monitor, and verify plans to remove these risks through reasonable means. SOX adopters are specifically obligated to conduct a formal risk assessment and choose a risk management strategy. Assessing risk is about identifying the strongest intersections between the likelihood of data compromise and its potential impact.

3) Managing Risk with Formal Controls

IT managers, chief data officers, and data governance types are starting to understand the language of "Controls," which result in physical steps and software to prevent, detect, correct problems around data at risk in this case. Preventive controls like a door lock are best, as are tools like a database firewall which do multiple things like monitor, alert, block, and audit adverse activity. IRI software prevents and corrects detected problems with data through data discovery, masking and auditing. Automating these types of controls is preferable because people make inconsistent mistakes.

4) Executives Are Responsible for the Controls

Because information in financial reports – like those behind the WorldCom and Enron scandals and the consequent SOX legislation – can flow through IT department resources, CEOs, CFOs and CISOs need IT's help to attest to the accuracy of the data and the efficacy of the controls. Although SOX requires this for financial reporting, consider also the need to comply with data privacy laws.

GDPR, HIPAA, PCI DSS, and many newer regulations like the CCPA, means that, among other provisions, you must guarantee data could not have been breached, changed, or hacked, and that you must be able to show how your company protects and provides the data. A good control system must also have logging built-in for verifying the use of these controls. Companies who safeguard PII and prove it, not only prevent fines and lawsuits, but enjoy the repeat business of more trusting customers.

5) Deploy Industry-Standard Controls

Using approved protocols for data and access protection supports compliance. Examples include:

- FIPS 140-2 and NSA Suite B encryption techniques

- Committee of Sponsoring Organizations of the Treadway Commission (COSO)
  Principles, reports, and recommendations from a private-sector consortium in the USA founded in 1985 "dedicated to improving the quality of financial reporting through business ethics, effective internal controls, and corporate governance."

- Control Objectives for Information and Related Technologies (COBIT)
  "IT governance framework and supporting toolset that allows managers to bridge the gap between control requirements, technical issues and business risks."

- International Standards Organization (ISO) 17799
  10 controls first published in 2000 comprising the best practices in securing information assets.

- Information Technology Infrastructure Library (ITIL)
  Published by the Office of Government Commerce in Great Britain, ITIL focuses on IT services and is often used to complement the COBIT framework.

- SANS Institute / Center for Internet Security - Critical Security Controls
  20 highest-priority recommendations for cyber defense updated since 2008 by an international, grass-roots consortium of corporate, government and institutional cyber analysts, hackers, solution providers, users, consultants, policy-makers, executives, academia, auditors, etc.

6) Implement Data Governance through People and Rules

Applications are easy to map to corporate organizational charts when a specific person or group is responsible for it. Data, on the other hand, flows throughout the organization and does not map to specific points on an organization chart. Only now are companies beginning to understand that they cannot map data accountabilities to specific job functions or departments.

The Data Governance Institute asserts that responsibility must be mapped across the organization. This is the goal of those in the "Data Governance Office." Data Governance sets the rules of engagement for people responsible for specifying, designing, implementing, monitoring, testing, and retiring the controls on data. The controls have a life cycle, just like applications and data do.

Data Governance officials are responsible for implementing these rules throughout the life cycle, and corporate IT staff provides input into, and is accountable to, these officials and the agreed-upon rules. It is up to data governance efforts to identify the location and nature (e.g., risk level, and need-to-know classifications) of data at risk.

7) Master Data Management (MDM)

Master data is a unique, core set of transactional data elements, or fields, used in many information systems' processes and transactions; i.e. control tables, or reference data in a lookup table. It exists because companies need to share data across departments and business functions, or, in the case of a merger or joint venture, two or more companies must share data across different platforms.

Master data is used for modeling and analytics, as well as in data governance. For example, it can be used in data quality rules used in data migration, cleansing, or in stewardship procedures, like those used to prevent or protect disclosures.

When a piece of master data that affects many applications and transactions is missing, wrong, inconsistent, or mislabeled, there will be adverse effects downstream. Therefore, good master data management (MDM) is required to keep this data clean, and to standardize master data models amid the relational taxonomy of facet data.

> *Ventana Research defines master data management as the practices and technologies allowing business and IT to define enterprise-wide master or reference data that is linked to the business. According to Ventana Global Research Director David Waddington, master data management enables companies to continue leveraging their current BI, ERP and data warehousing investments.*

Jane Griffin, a Deloitte Consulting partner specializing in data warehousing and business intelligence systems, posits in her article Information Strategy: Building a Data Management Strategy that there are four processes essential to a good MDM strategy:

1) Data Migration and Integration
2) Data Maintenance
3) Data Quality Assurance *and Control*
4) Data Archiving

Master field data is typically stored in a centralized repository – often in support of a Service-Oriented Architecture for cleansing and sharing – suggesting a higher stakes need for field-level protection. MDM is thus also a data governance issue, and there is a need for it to be protected by the same logic and methods as transactional data.

For more information from IRI about this issue, please refer to:
https://www.iri.com/solutions/metadata-mdm/master-data-quality-and-security

Startpoint Security

As mentioned in the introduction, IRI refers to the processes around data-centric protection or data masking within a newly defined rubric of "Startpoint Security." IRI SVP and COO David Friedland coined this term in 2018 to better organize the concepts that provide the context and justification for data masking.

Background

When it comes to PII in the various points along modern networks, "*endpoint* security" tools and techniques first come to mind. They have been the most obvious and advertised way to protect the PII stored in mobile phones or provided through apps, as well as laptops, desktop PCs, servers, and networks/switches through which they connect. We also think about storage devices like thumb and hard drives, or folders, files, and entire databases that can be encrypted within them.

Fortunately, the IT industry also thinks about securing this data as it enters and moves around transaction and analytic systems. Within IRI's new definition of "*startpoint* security" are several targeted ways to find, secure, and account for the PII in the databases and files that applications use. After all, it is in those repositories where PII is first created for storage and queries, integration and analytics, and ultimately moved out, protected or not, along the endpoints.

Indeed, IRI has focused historically on the granular protection of PII in data silos at rest (static data masking), and in transit (dynamic data masking). Securing PII directly in these sources or startpoints instead of just its endpoints continues to provide several benefits, including:

- Efficiency — It's much quicker, and less resource-intensive, to encrypt or apply other de-identification functions to discrete values than to everything else around them.
- Usability — By masking only what's sensitive, other data around it is still accessible.Those secured values can also move safely between databases, applications, and platforms on-premise or in the cloud.
- Breach nullification — Any misappropriated data is already de-identified.
- Accountability — Data lineage and audit logs pointing to specific element protections are a better way to verify compliance with privacy laws applicable to specific types of PII.
- Security — Some data masking techniques cannot be reversed, and many applied at once are harder to reverse than a single technique. Think as well about the difference in vulnerability here, too; i.e., if the encryption key is compromised, an entire network could be exposed instead of a single column of data.
- Reversibility — If the masked PII needs to be reversed, that is possible if a function like encryption or lookup-pseudonymization was used.
- Testing — Masked production data can also be used for prototyping DB and ETL operations, platform benchmarking and DevOps.

The actual de-identification of the data is nonetheless only one piece of a larger approach.

Definitionrefere

At first, it would seem that startpoint security is just another IT industry buzzword synonymous with data masking. However, masking is just one step in the broader industry context of risk and compliance.

Under IRI's definition, startpoint security *also* takes into consideration eight more related pieces:

1. Permission & Disclosure — authorizing you to store submitted PII via user agreement
2. IAM & RBAC — managing access to data sources, (un)masking jobs, and programs
3. Discovery & Classification — searching and cataloging PII to find and mask it consistently
4. Data & Metadata Lineage — saving and analyzing changes to data and masking jobs
5. Latency — architecting and configuring static or dynamic data masking jobs
6. Risk Scoring — determining the statistical likelihood of re-identification (for HIPAA EDM)
7. Audit Logs — seeing or querying who did what, and who saw what, when, and where
8. Assessment & Insurance — conducting expert procedural, statistical, and legal reviews

Many of these additional considerations are not exclusive to startpoint security, but data classification, lineage, latency, and re-ID risk measurement are certainly more relevant in the data-centric realm than they are to endpoint security.

Application

Data masking products and practices are typically deployed in response to a breach or to help comply with a particular privacy law. Most data masking tools usually provide just one or two of the capabilities listed above, and analysts recognize such vendors as narrowly.

IRI has endeavored to offer more in this space by addressing these well-defined concerns, and combining them in each of the static and dynamic data masking tools in the IRI Data Protector Suite.

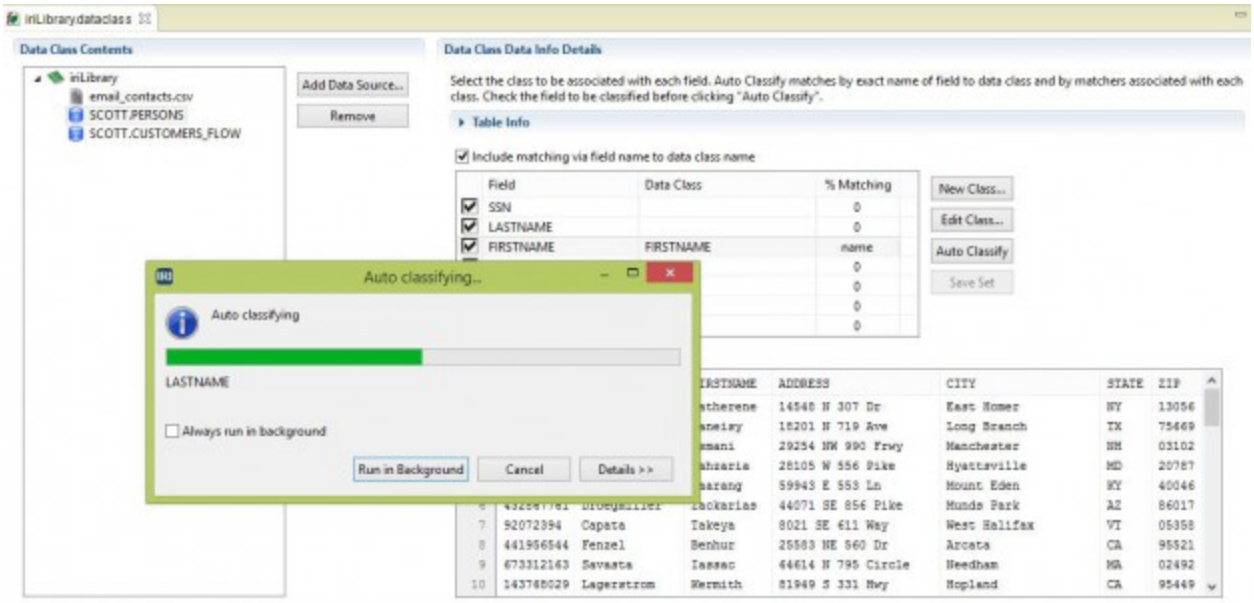
IRI also includes all of its static data masking and test data generation tools within its larger Voracity data management platform for data discovery, integration, migration, governance, and analytics. In Voracity, all jobs are built and run in the same Eclipse IDE, are powered through a common single-pass data mapping and masking engine, and use the same data definition and manipulation metadata.

The following recommendations cover a few of the startpoint security elements listed above, as well as other concepts less often considered, but equally important.

General Recommendations

1) Classify & Discover Data at Risk (for Deterministic Masking & Referential Integrity)

To mask PII, you must first define and find it. According to Gartner, only half of companies needing to mask PII know precisely where it is. IRI data classification, profiling and searching tools group and locate PII according to defined pattern-, lookup value-, and fuzzy- matched attributes, or based on machine-learned NER models. Once that data is classified, discovered and validated, you can apply the same masking functions to it -- making ciphertext consistent, and preserving referential integrity.



2) Structure & Standardize the Data

Flat records – including those in COBOL, CSV and LDIF files – and other interchange formats like HL7, XML and JSON, can be a good common denominator for an organization's sources of truth. Mainframe data, and data on the way to and from databases and other applications like spreadsheets, reporting tools, and web logs often reside in flat files. Many of them contain PII which can be quickly and differentially masked, and then re-targeted into DB silos, reports, apps, etc.

Big data stored in flat files is easier to define, compress, store, process, report from, and protect than data in tables or web service applications. On secure servers, the files can be rapidly masked on a standalone basis, or in the course of data warehouse extract, transform, and load (ETL) jobs. Big data processing tools like the SortCL program in IRI CoSort run in the file system, bypassing the overhead of slower DBMS I/O and SQL encryption functions, and performing multiple transformations at once. Data in HDFS can also be masked through interchangeable Hadoop engines in Voracity without re-coding.

If the data are not in structured sources, consider flattening it into structured flat files. The dark data discovery tool in IRI Workbench can pattern-search, extract, and produce delimited files which can then be masked and re-targeted. Alternatively, IRI DarkShield can find and mask that discovered data at the same time, or afterwards, exactly where it resides in log, text, MS Office, Parquet, pdf and image files.

3) Embed Masking Functions

There are many products designed to hide, obfuscate, and protect data, including file and disk encryption hardware and software. And, there are many ways to use and prepare data for analytics in integration environments like the data lake, corporate information factory (CIF), operational data store (ODS), and data warehouse (ETL) environments. But the two activities rarely meet. Yet when:

- data volumes are growing
- production windows are shrinking
- resources are finite

data masking functions should be applied *in the same program and I/O pass* with data cleansing, transformation, migration, and reformatting and reporting jobs to save time and money designing and running . This can be done by adding FieldShield masking functions to your data targets in SortCL job scripts created to run in IRI CoSort product or IRI Voracity data management platform workflows.

4) Protect Master Data

Risk is centralized in a single place when users move master data to a central repository where it is easier to clean and supports SOA. Consider also the need-to-know and encryption requirements in the case of outsourcing and testing when the master data is "all in one basket." Special security filters may thus be needed when storing and moving those values, too.

Static masking jobs, as well as authenticated dynamic DB masking jobs, in IRI FieldShield, can enforce need-to-know rules on master data through their user-specific security techniques.

5) Use Safe Test Data

Test data has many uses, including:

- AppDev / DevOps
- Range & Stress Testing
- User Acceptance Testing (UAT)
- Format Sharing or Outsourcing
- Database Simulation
- Platform Benchmarking

Common techniques for the generation of useful test data have included custom programs, shareware tools for column generation, and taking selected snippets of production data. This last method provides the true essence of the field elements and the actual look and feel of the file or table containing them. However, production data is an inherently unsafe, if not prohibited, source of test data.

Data masking software is commonly used to morph production data to make it safe for disclosure and useful for testing, particularly when realistic functions like pseudonymization and format-preserving encryption are used. However, production data may not always suffice; i.e., it may not yet exist, cannot be accessed, or does not meet future application or stress-testing criteria.

In such cases, a fit-for-purpose test data synthesis facility is needed. IRI RowGen is a product and test data method involving random generation of data based on the metadata of production tables or files. It can rapidly populate many realistic value ranges, target structures, and dependent key relationships.

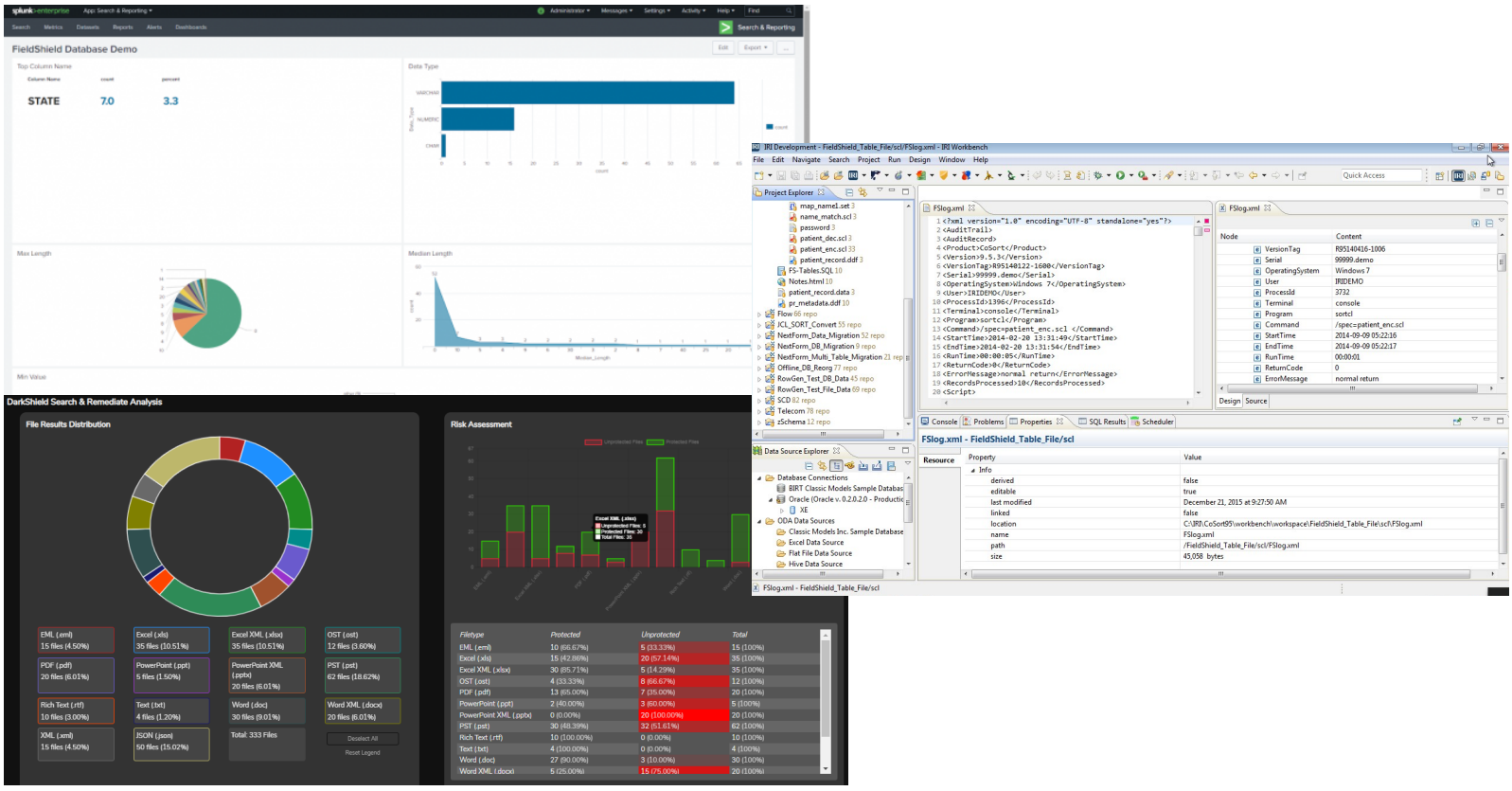
IRI data masking, subsetting (and masking) and synthesis operations can run standalone, or from the IRI Workbench front-end job design GUI, built on Eclipse. These same operations can also be invoked from IRI APIs or third-party DB cloning and virtualization platforms like Actifio, Commvault and Windocks, or the Value Labs Test Data Hub and Cigniti test data management web applications.

6) Verify Compliance

Keep performance and transaction records of data processing, reporting, protection and generation steps. Statistics should be kept in secured, query-ready audit logs that include:

- Runtime information; e.g., when the job was run and how long it took
- Application statistics; e.g., number of records read, transformed, masked, etc.
- Source and target names, including those resolved from environment variables
- All field specifications; e.g., name, position, size and data type attributes
- All field and file manipulations by showing the actual job script used
- Re-ID risk scores from key and quasi-identifiers using approved statistical methods

IRI data masking tools create such trails to report the actions taken and verify compliance, as well as facilitate remediation internally or externally via SIEM tools like Splunk or “playbooks” like Phantom.



IRI Workbench data profiling results in a Splunk ES cloud dashboard, an XML audit log entry for an IRI FieldShield data masking job, and IRI DarkShield unstructured search and remediation results.

“Safe Data” Techniques (Masking Functions)

Column, field, or even more granular value-level data masking allows data governance officials and IT staff to produce data views or persistent targets that can leave the office or firewall without compromising PII. Masking data is also faster and more useful than hiding whole rows, files, tables, DBs, etc. because the masked data can still be used, and concealed or revealed in very specific ways.

These techniques help, for example: healthcare companies to achieve compliance with HIPAA regulations by de-identifying or shrouding medical and personal data; to satisfy government agency and contractor needs to redact sensitive values using various techniques; and, to allow payment card processors to adhere to PCI standards via the encryption of credit card numbers.

Data Omission & Deletion

Select which input rows, values or columns will go to target tables, reports, and hand-off files on a need-to-know basis, based on your business rules and privacy laws. This process can start with data classification, where only defined items or groups are selected for global discovery and remediation.

However, even without a discovery or broader process like classification, individual masking jobs should be able to specify conditional criteria for withholding creation or display of data. Such conditions should be applicable at the row, column, or value level, just as they should allow you to apply masking functions on the data retained, and to delete (erase) PII values to support the ‘Right to be Forgotten.”

Anonymization

Remove the individualizing characteristics of data so that a person or item stored in the original field cannot be identified. Once anonymized, the data cannot be linked to any source. The benefit of anonymization over filtering or non-format-preserving encryption is that the original field layout – position, size, and data type – can remain the same, and thus remain realistic for test or demo use.

There are several ways IRI software users can anonymize data, including:

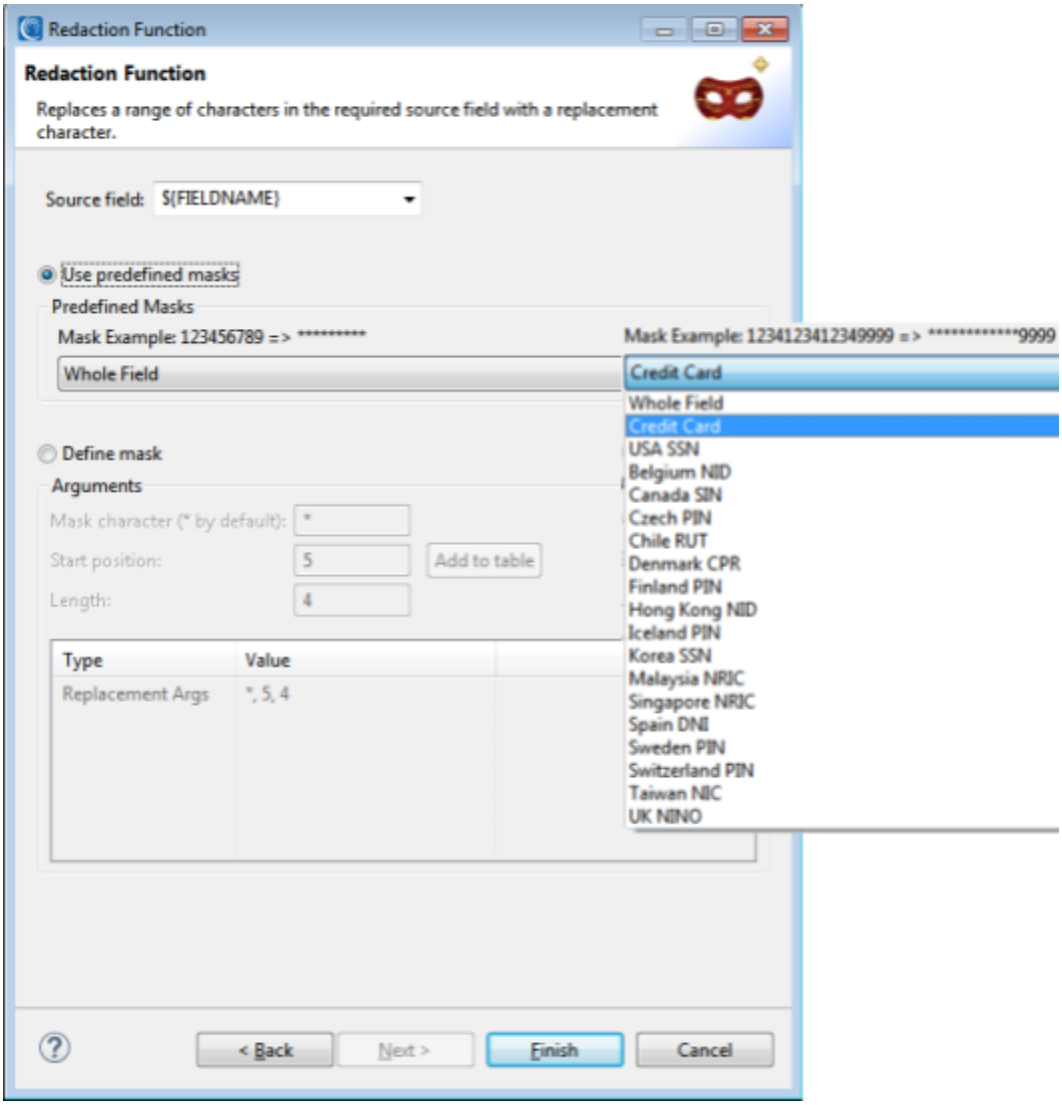
- blurring ages, dates and other numeric values with random noise
- bucketing or generalizing quasi-identifying attributes like age, occupation, marital status, or race
- using mathematical expressions on numeric data
- character shifting or bit manipulation (scrambling)
- literal replacement or string functions
- hashing
- tokenization

IRI currently supports fifteen categories of data masking functions for application at the row, column, or value level. The three most popular are redaction, encryption, and pseudonymization. The method chosen depends on factors like strength, ciphertext appearance, reversibility, uniqueness, and speed.

Redaction / String Masking (Non-Recoverable)

Data redaction, or string masking, is a form of irreversible de-identification that involves obfuscating one or more bytes of a value with chosen characters or black-outs, permanently hiding the original value.

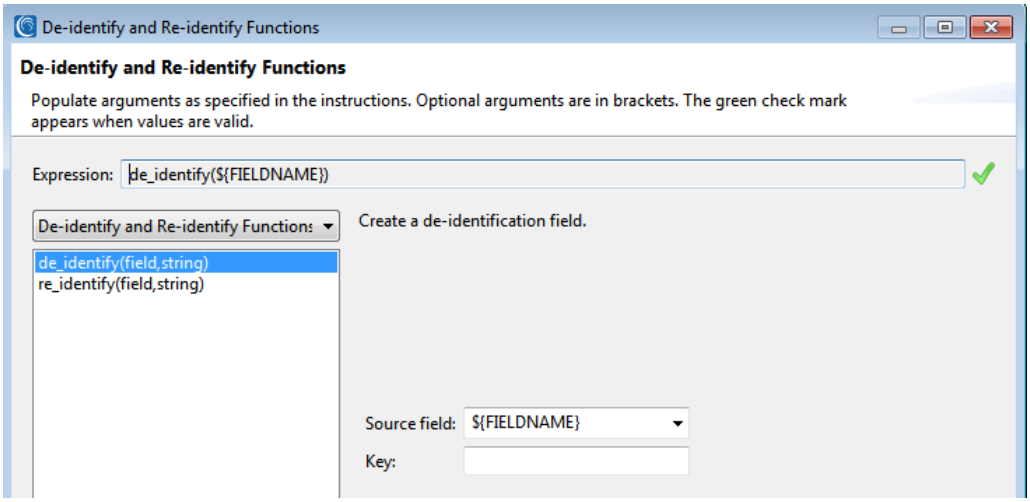
This can be applied to the data based on conditions, so that only those values meeting the specified criteria logic are redacted.



Data Redaction function dialog in the IRI Workbench IDE for FieldShield, DarkShield, etc, built on Eclipse™.

Bit Scrambling (Recoverable)

Change the individualizing characteristics of data so that a person or item stored in the original field cannot be identified, but nonetheless remain individualized so that it can be followed safely through different departments, and then if necessary, re-identified. This method is faster but less secure.



De- and re-identification dialog for ASCII values in IRI Workbench

Encryption and Decryption (Recoverable)

The encryption of data in files, databases, and on disks and other media has been practiced for years. According to expert James C. Foster, author of Look Before Leaping into Database Encryption:

> *Encryption is a powerful security tool, and nearly every compliance standard or industry regulation addresses data security in some manner, often at least implying a role for encryption. For instance, the Gramm-Leach-Bliley Act (GLBA) requires organizations must "insure the security and confidentiality of customer records and information," and California's SB 1386 breach-notification law states that any breach of the security of unencrypted personal information must be disclosed.*
>
> *Here are some simple guidelines that will help you secure your database without impeding the business you're trying to protect:*
>
> - *Never encrypt foreign or super keys (encrypted keys used for indexing could cause usage and performance issues).*
> - *Use symmetric over asymmetric cryptography when available (again, for performance).*
> - *Full database encryption is rarely advised or a feasible option. Security best practices would teach you to encrypt everything with multiple keys and differing algorithms. However, the significant performance hit you must selectively choose.*
> - ***Encrypt only sensitive data columns. This is typically all that is required or recommended by regulations and, after all, is what needs protection.***

Determining key fields or data elements is a daunting task and should be driven by compliance and threat mitigation. Since no regulation comes right out and states "X" columns must be encrypted, it falls back on good judgment. Identifying your most sensitive data and all the places it may reside, from primary databases to backups, is one of the toughest parts of implementing encryption, which is why companies may attempt to solve the problem by deciding to simply, if misguidedly, "encrypt everything."

Choosing an appropriate encryption method also depends on your data. If it mainly consists of images and Web content, then a weaker algorithm – such as DES or SSL – may be adequate. However, if you are storing personally identifiable customer information or the company design for a nuclear disintegrator, choose strong encryption with a larger key space, such as AES, Blowfish or 3DES.

The choice of encryption products is improving, as database encryption has made significant strides in the past few years, and the market has continued to mature.
…

While compliance pressures have stoked keen interest in database encryption, it doesn't solve all database security concerns, which are at least as much about preventing abuse by privileged insiders as external attackers. There are no shortcuts. Hastily implementing database encryption simply to comply or assuming it alone will make your data secure will cost extra time, money, manpower and brainpower better spent elsewhere.

Encryption is the most secure of the field protection techniques presented. It uses technology that conforms to US government standards, like FIPS, for hiding the contents of sensitive data. On output, the fields are in a largely unusable form until decryption with the proper key occurs.

Field-level encryption is worth considering because of its flexibility and performance benefits:

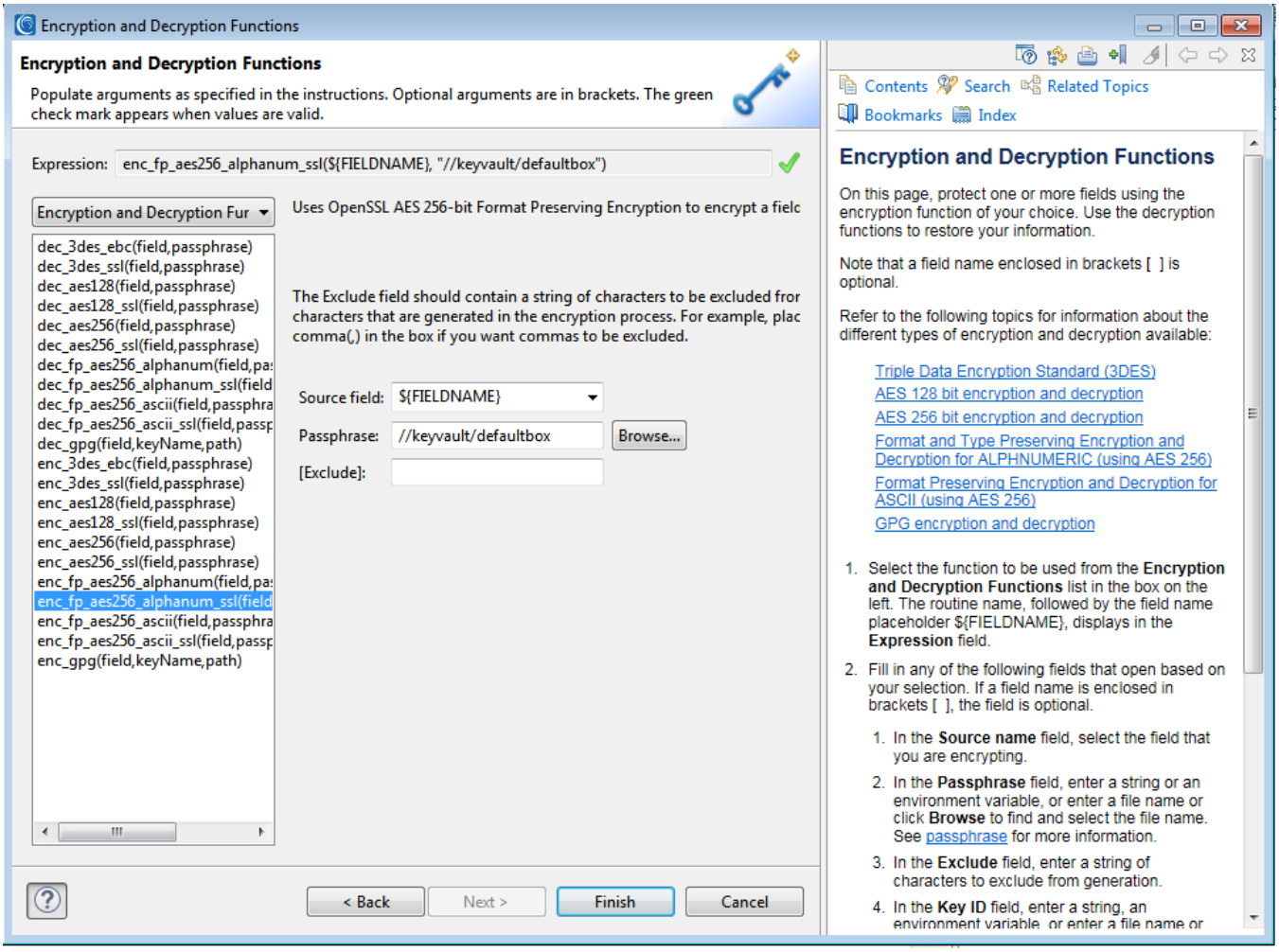
- Only sensitive fields are encrypted; remaining fields and disk are readable
- Different encryption keys and libraries can be used on different fields
- Common encryption functions and keys supports referential integrity
- Different field protection methods can be used simultaneously
- Encryption can occur during routine data transformation and reporting
- Computing resource overhead is nominal

IRI offers several choices for encrypting fields in multiple data sources, including database columns, flat files, etc. FieldShield, CellShield, DarkShield or Voracity users can encrypt with functions like format-preserving AES-256, AES-128, FIPS-compliant OpenSSL, 3DES, and GPG, or their own runtime-linked functions.

IRI's Suite B-compliant AES implementations can be managed through with user-provided (encrypted) key passphrases, files, environment variables, or key vaults (e.g., in Azure or Townsend Alliance Key Manager) that contain encryption keys. The passphrase acts as a seed to a function that generates a Secure Hash Algorithm hash digest to derive the encryption key.

Consider also that:

1) Encryption key management is currently the best way to define and manage role based access control (RBAC) at very granular levels; i.e., for specific classes of data, columns or cell ranges.
2) Database vendor encryption products provide fewer functional choices, can be cumbersome and expensive to implement, and cannot be used on other databases or flat files.



FieldShield encryption and decryption dialog in IRI Workbench

By way of application, consider the Final HIPAA Security Rule enacted in 2003 governing the protection of electronic private health information (EPHI):

> *Section 164.312, **Technical Safeguards**, contains provisions extracted from two sections of the proposed rule: Technical Security Services and Technical Security Mechanisms. **Covered entities must implement:***
>
> ***Technical policies and procedures for access control on systems that maintain EPHI. These systems must allow for unique user identification** and include an emergency access procedure for obtaining necessary EPHI during an emergency. Addressable specifications include automatic logoff and encryption and decryption, which is defined as **a mechanism to encrypt and decrypt EPHI**.*

With control at the field level, multiple encryption libraries and passphrases can be used for field-specific need-to-know decryption entitlements.

- *Transmission security, including two addressable specifications:*

  1. Integrity controls – security measures to ensure that electronically-transmitted PHI is not improperly modified without detection until disposed of, and

  2. Encryption. Designation of encryption as an addressable specification is a key departure from the proposed rule, which explicitly required encryption when using open networks. Covered entities now must determine how to protect EPHI "in a manner commensurate with the associated risk." ***Covered entities are encouraged in the Rule's preamble to consider the use of encryption technology for transmitting EPHI, particularly over the Internet.*** *The key reasons cited by HHS for this change are the cost burden for small providers and the current lack of a simple and interoperable solution for email encryption.*

IRI makes low-cost encryption another option, along with many other anonymization and masking functions at the field level, while running routine data manipulations and reports.

- *Hardware, software, and/or procedural methods for providing audit controls.*

FieldShield audit records include the full job script, along with the path and name of the encryption libraries. The secure audit log can be used to query and display what, when, how, and by whom the PHI field data was encrypted, or and otherwise masked or transformed.

- *Policies and procedures to protect EPHI from improper alteration or destruction to ensure data integrity. This integrity standard is coupled with one addressable implementation specification for a mechanism to corroborate that EPHI has not been altered or destroyed in an unauthorized manner.*

Data that does not decrypt with the proper encryption key suggests that the decrypted field has been compromised. This can be traced in logs that track when the file was processed for field encryption.

- *Person or entity authentication, which requires the covered entity to implement procedures that verify that a person or entity seeking access to EPHI is the one claimed to be doing so.*

Passphrases are used inside IRI data masking software to generate keys for encryption and decryption of field data. Therefore, only the person or entity in possession of the right libraries and passphrase(s) can encrypt or decrypt data.

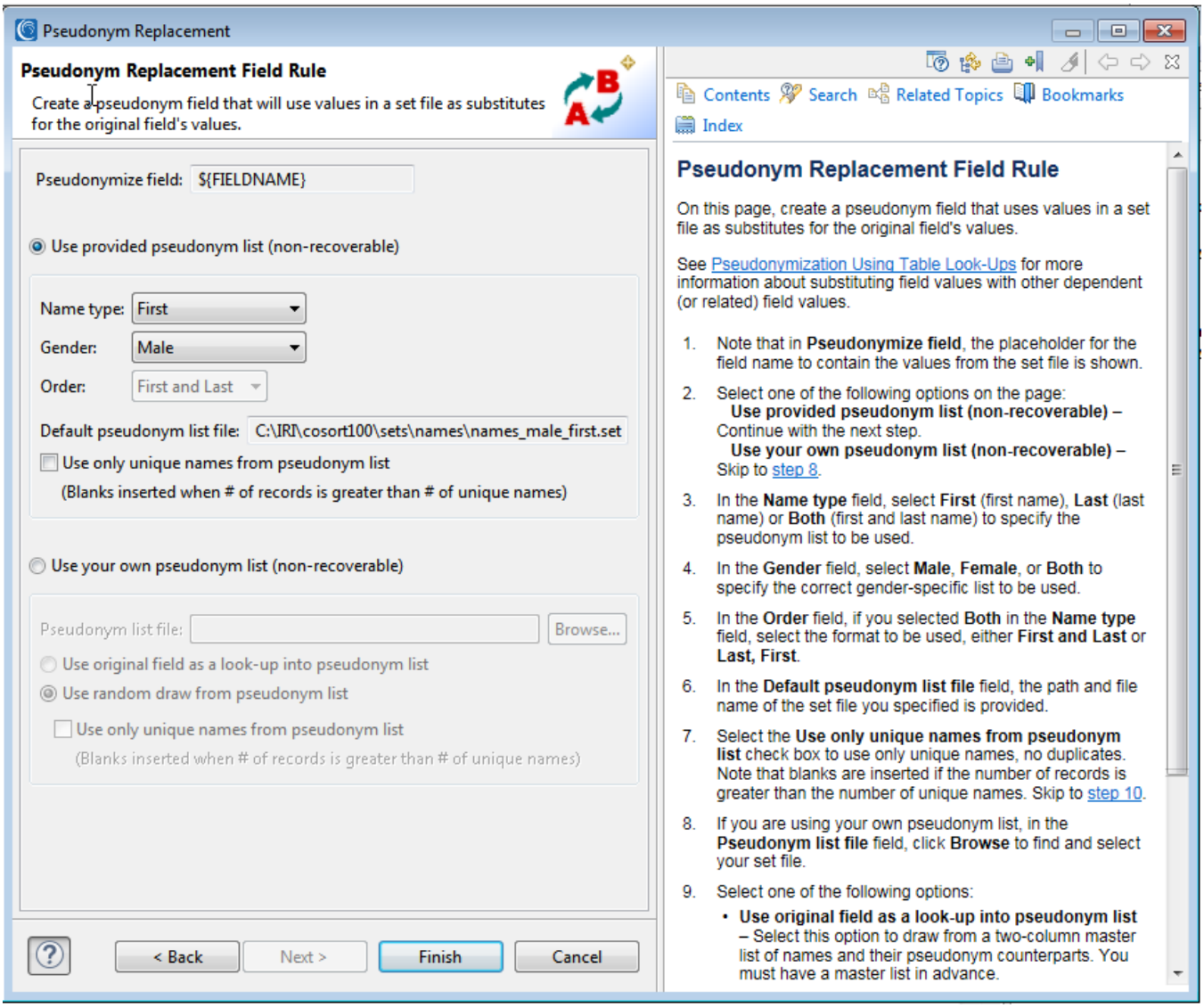
The passphrase can exist inside the job script, implicitly through an environment variable to hide the passphrase, stored in a file within a permissions-restricted directory, or connected and rotated through a third-party key store or HSM, like those in MS Azure Key Vault or Townsend Alliance Key Manager.

Pseudonymization

One of the best ways to protect the most identifying piece of personal information – someone's name – is to use a fake name, or pseudonym. IRI software provides two methods for replacing the original value of a field with a substitute value:

Non-Recoverable Pseudonymization

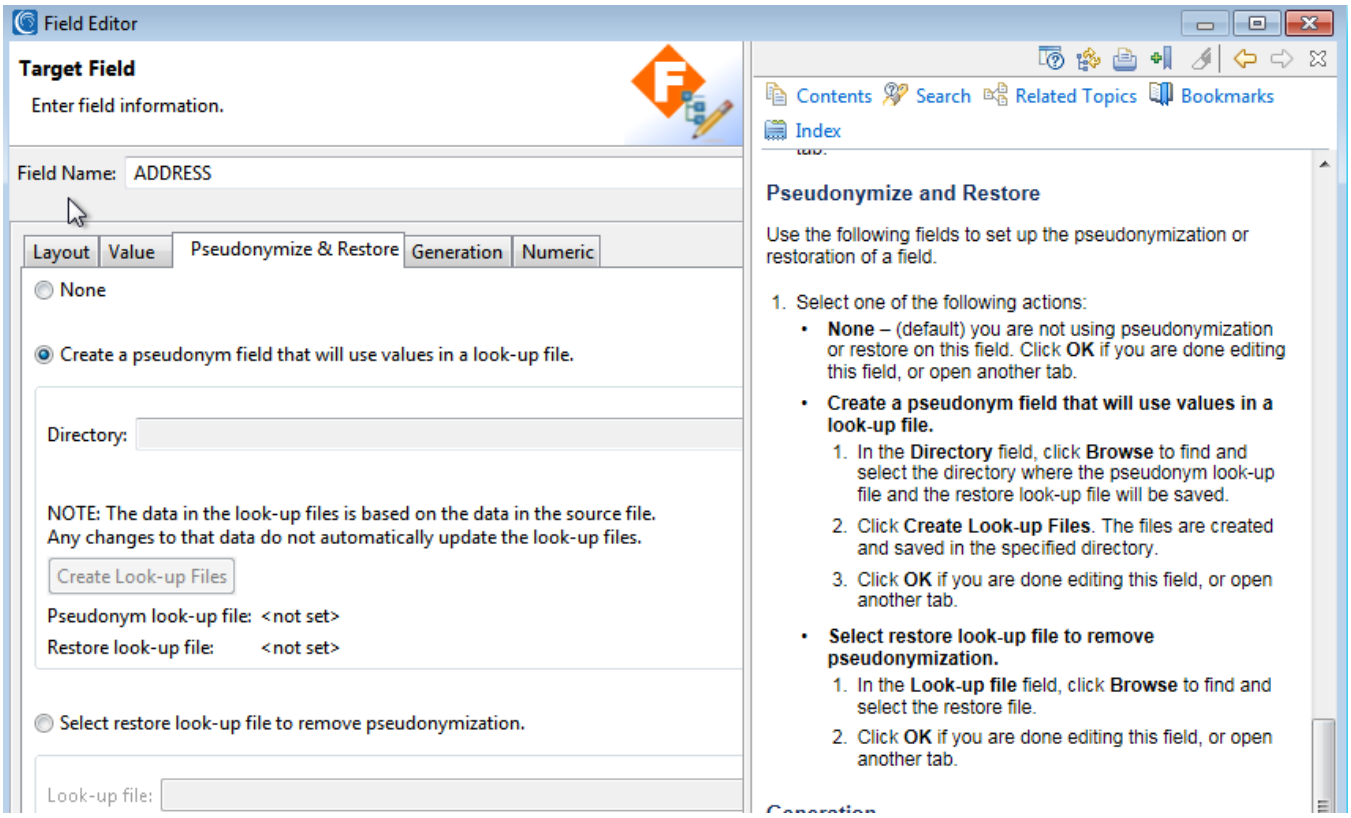
This function performs a random lookup from a supplied list, or single-column set, file of names. IRI supplies gender-specific western first, last or both names, though any other set can be used. This allows names to appear real without compromising actual identities. This approach does not allow the original names to be restored because there is no matching value association.



Pseudonym value creation and specification dialog in IRI Workbench

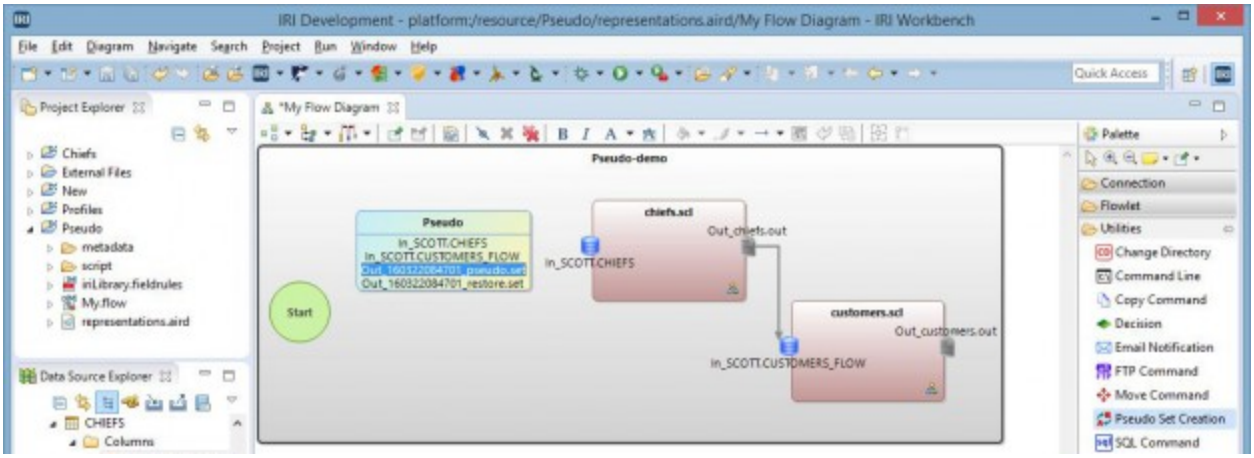
Recoverable Pseudonymization

This function uses or builds a sorted lookup, or two-column set, file containing the original names from the input source, and substitution values randomly selected from the lookup set. This creates a shuffled list of real names so there is no direct association between the original people and their attributes within other columns. An inverse set file can also be built, so that recovery is possible in the same manner.



Pseudonymization Restoration dialog in IRI Workbench

Additional consideration must be given to pseudonymization when changes to sources and multiple database tables are involved, so that uniqueness and referential integrity can still be preserved.



Randomization (Non-Recoverable)

Replacing real values with random values is yet another approach to shielding PII from disclosure while maintaining the original structure of the input sources. This process is not reversible and thus cannot be used to maintain referential integrity. IRI provides two methods for replacing the original value of a field with a substitute value:

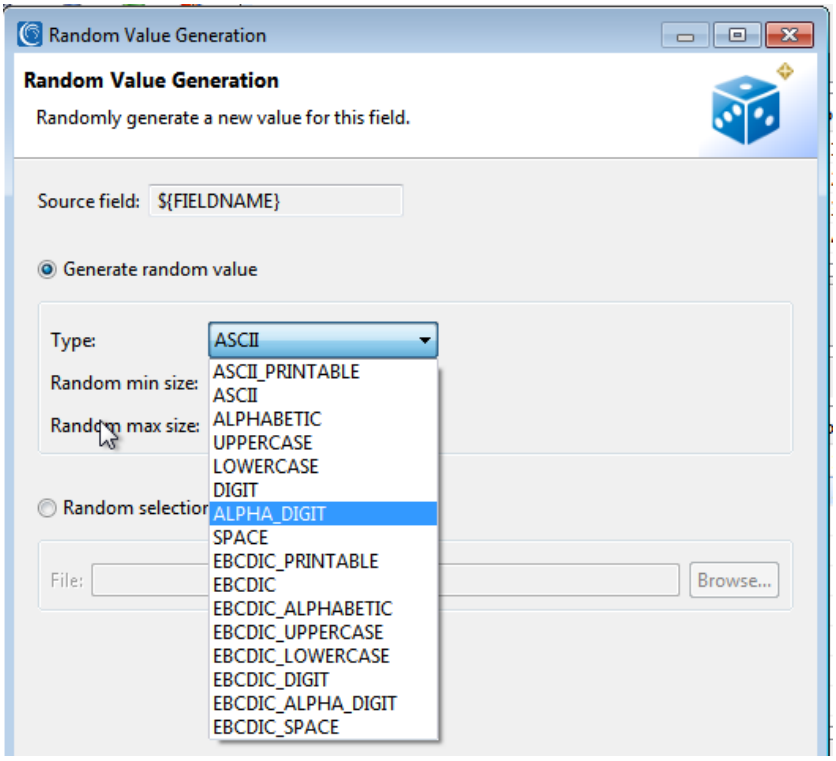
Random value generation

Source field values are replaced with randomly generated values of a given data type. Using random data, especially numbers, can make the protected field look real, while not requiring the use of any real data or protective overhead. See these additional data generation rules available to IRI software users.

Random value selection

Source field values are replaced with randomly selected values in a lookup, or "set", file containing:

- an alphanumeric list of values: the values appear in output exactly as they appear in the set file, producing realistic-looking data and the values that appear in output contain only the numbers, or a specified range of numbers, in the set file
- date values and ranges: only listed dates, or valid dates within a specific range, appear



Random value creation dialog in IRI Workbench

Conclusion

Your organization manages personally identifiable information (PII). Your data governance efforts must prevent the kinds of data disasters posted at the Privacy Rights Clearinghouse. You must comply with industry and government data privacy rules, to keep data safe in production and test environments.

Where you cannot eliminate PII, you have to define/classify it, find it, de-identify it, and verify that you protected it. Then you have to continue monitoring and addressing data risks going forward.

At protection time, technology choices are difficult. Traditional encryption of entire databases, files, disks, or devices is inefficient, especially in volume. It also restricts access to non-sensitive data, and is subject to complete exposure from a single password breach.

Meanwhile, certain data masking methods and products produce insecure results and may not work in your environment; i.e., they render the protected data unsuitable for testing, marketing, or research purposes. Moreover, with current methods you may not get:

- an audit trail detailing how you managed risk – forcing costly validation exercises
- the ability to simultaneously apply multiple, or consistent, protections to multiple sources
- the ability to combine data protection with other data processing operations
- a single, affordable solution for different data sources and hardware platforms

It is therefore important to understand the value of data masking operations like encryption, per:

> *"Solutions that provide encryption at the file, database field, and application level provide the highest level of security while allowing authorized individuals ready access to the information. Decentralized encryption and decryption provide higher performance and require less network bandwidth, increase availability by eliminating points of failure, and ensure superior protection by moving data around more frequently but securely."*
>
> *– Gary Palgon, Enterprise Systems Journal*

and to do so within the larger context of startpoint security.

A well-designed set of processes and tools should be used to secure PII in ways that render it fit-for-purpose, safe from breaches, and compliant with data privacy regulations. This in turn supports the risk and controls framework of your enterprise as well as specific DLP and stewardship goals. It also promotes greater consumer confidence in your organization's ability to protect PII, and it facilitates seamless interoperability with other data management activities.

IRI Data Masking Technologies

Applicable data-centric security software products or services in the IRI Data Protector suite include:

IRI FieldShield	classifies, finds, masks and audits PII in structured file, database, or HDFS sources
IRI DarkShield	classifies, finds, masks and audits PII in structured, semi- and unstructured sources
IRI CellShield	classifies, finds, masks and audits PII in Excel sheets, local or LAN-wide
IRI RowGen	synthesizes or subsets realistic and referentially correct test data in RDBs and files
IRI Workbench	free Eclipse GUI for data profiling, database operations, and IRI job management

In addition:

IRI Ripcurrent	auto-applies FieldShield masking functions to DB rows incrementally, in real-time
IRI Voracity	includes everything above, plus subsetting, ETL, data quality, migration, analytics
IRI DMaaS	is a professional data masking service, including PII discovery and post-fix audits

See also this matrix, which drills down into more detail to help you select the right product.

These offerings support the profiling and protection of sensitive data, privacy law compliance, and testing through many static (SDM) or dynamic data masking (DDM) functions for PII in many sources.

FieldShield, for example, marries a familiar Eclipse GUI with a powerful metadata and a program, to:

- Classify, search, and profile PII with wizards built for multiple silos and schemas
- Pseudonymize, blur, scramble, encode, hash, randomize, and tokenize
- Encrypt and decrypt with multiple built-in (or your own) libraries
- Mask fields, rows, strings or faces based on rules or conditions
- Score re-ID risk per HIPAA EDM, FERPA,and GDPR rules
- Produce job-specific, query-ready XML audit logs

All the above products are also included in the 'total data management' platform called IRI Voracity. In addition, Voracity supports creation and masking of database subsets, plus any combination of techniques to produce fully customized test sets for an unlimited range of applications. However, Voracity is actually designed for more. It consolidates and accelerates the key enterprise data lifecycle management activities of data discovery, integration, migration, governance, and analytics.

FieldShield, RowGen and Voracity all use IRI Workbench, plus the same GUI and data definition and manipulation metadata with other IRI tools, including IRI CoSort for data transformation, cleansing, and reporting and IRI NextForm for data and DB migration and replication. This overlap makes it easy to add data masking to those jobs – *plus* any other enterprise information management (EIM) jobs that Voracity supports, including DW ETL, BI/analytics, DevOps, etc.

In summary, data masking software and services from IRI or experts you choose – can find and secure sensitive data, big or small, in your charge, for bespoke data security requirements or as part of a larger data governance and (test) data management framework.

Total Data Management

Innovative Routines International, Inc.
2194 Highway A1A, Third Floor
Melbourne, Florida 32937 USA
Tel. +1.321.777.8889
https://www.iri.com

info@iri.com