



**IRI DarkShield**  
Unstructured Data Search & Security



***Version 4***

# **Product Overview**



## Executive Summary

In an information technology era in which both big data opportunities and privacy law exist and converge, there is a pressing need to discover, work with, and protect data hidden in unstructured files. Data in these repositories is often referred to as dark data:

*Gartner defines **dark data** as the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing). Similar to dark matter in physics, dark data often comprises most organizations' universe of information assets. Thus, organizations often retain dark data for compliance purposes only. Storing and securing data typically incurs more expense (and sometimes greater risk) than value.*

Every company and government agency collects and stores such data in logs, email repositories, documents, images, and audio/video format. Like transactional data in structured sources, the information contained in unstructured data sources carries both analytic value and business risk.

Innovative Routines International ([IRI](#)), Inc., founded 1978 and best known worldwide as The CoSort Company, expanded its high-volume, high-performance data transformation capabilities into the world of sensitive data discovery and masking in 2007. The addition of encryption, redaction, pseudonymization, and other anonymization functions was a natural evolution of the field-level manipulations IRI software was already performing in mainframe sort and data migrations, big data integration and wrangling, test data generation, custom reporting, and so on.

IRI has created fit-for-purpose data masking tools from this foundation, and has enjoyed both commercial success from them, and recognition from the data security analyst community; e.g., Gartner, which now features five IRI products in its [Market Guide for Data Masking Technologies](#). IRI's latest offering, DarkShield®, is designed to reduce the cost and risk involved in finding and securing information in dark data repositories, and to help you nullify the risk of data breaches and comply with data privacy laws.

For its innovations in PII security for unstructured data in relational and NoSQL DBS, DarkShield was recently recognized as a [trend-setting product](#) by DBTA Magazine.

## Contact Information

Innovative Routines International, Inc.  
2194 Highway A1A, Suite 303  
Melbourne, FL 32937 USA  
Tel. +1.321.777.8889  
[darkshield@iri.com](mailto:darkshield@iri.com)



## Product Introduction

[IRI DarkShield](#) Version 4 is a software package for finding and masking Personally Identifiable Information (PII) and other sensitive data hidden within semi-structured and unstructured files, as well as relational and NoSQL databases. It can be licensed and used standalone, or within the [IRI Voracity](#) data management platform.

DarkShield utilizes any combination of regular expressions, value lookups, path filters, and Named Entity Recognition (NER) models to search for PII floating in Microsoft documents, as well as .pdf, json, and other plain text files like logs and emails, plus most image file formats through Optical Character Recognition (OCR). It also supports bounding box areas and facial recognition to redact sensitive positional PII in image files. Support for compressed and A/V files, more proprietary formats, and direct connectors to cloud and other silos is now in development.

In the same or separate pass from the search operation, DarkShield can extract for delivery (data portability), mask with industry-standard protection functions, and report on the found values and file metadata they are in. Supported masking functions include: redaction, encryption, pseudonymization, hashing, encoding, bit and string maneuvers.

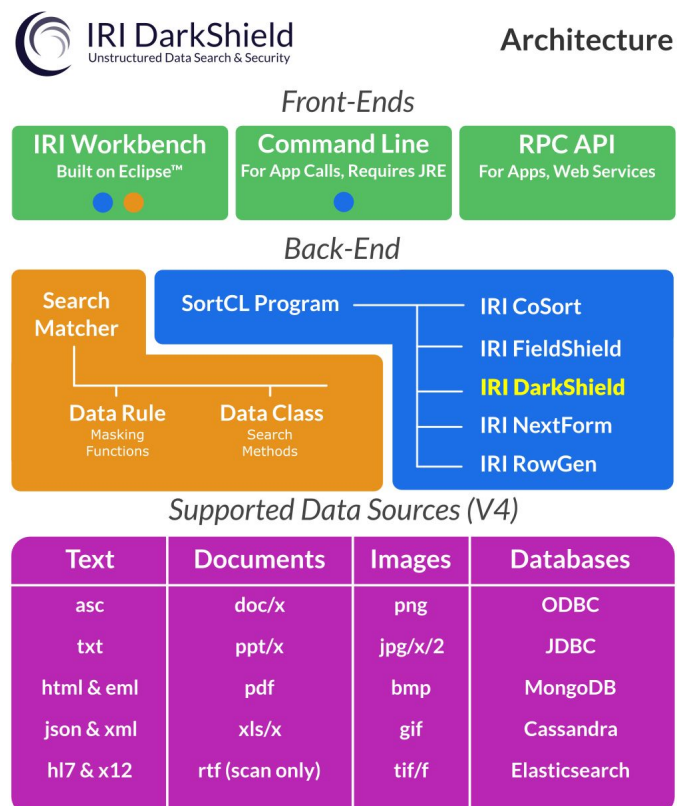
DarkShield-process metadata is serialized in XML for easy modification and repeat use, and can be shared in cloud repositories like Git. Search and masking results are in a flat-file log that can be audited in several ways, including: ad hoc, via built-in interactive dashboard, CoSort SortCL query programs, or Datadog and Splunk (ES, etc.) analytics.

## DarkShield Architecture

DarkShield search and mask operations are powered by IRI CoSort and other proven big data and data science technologies.

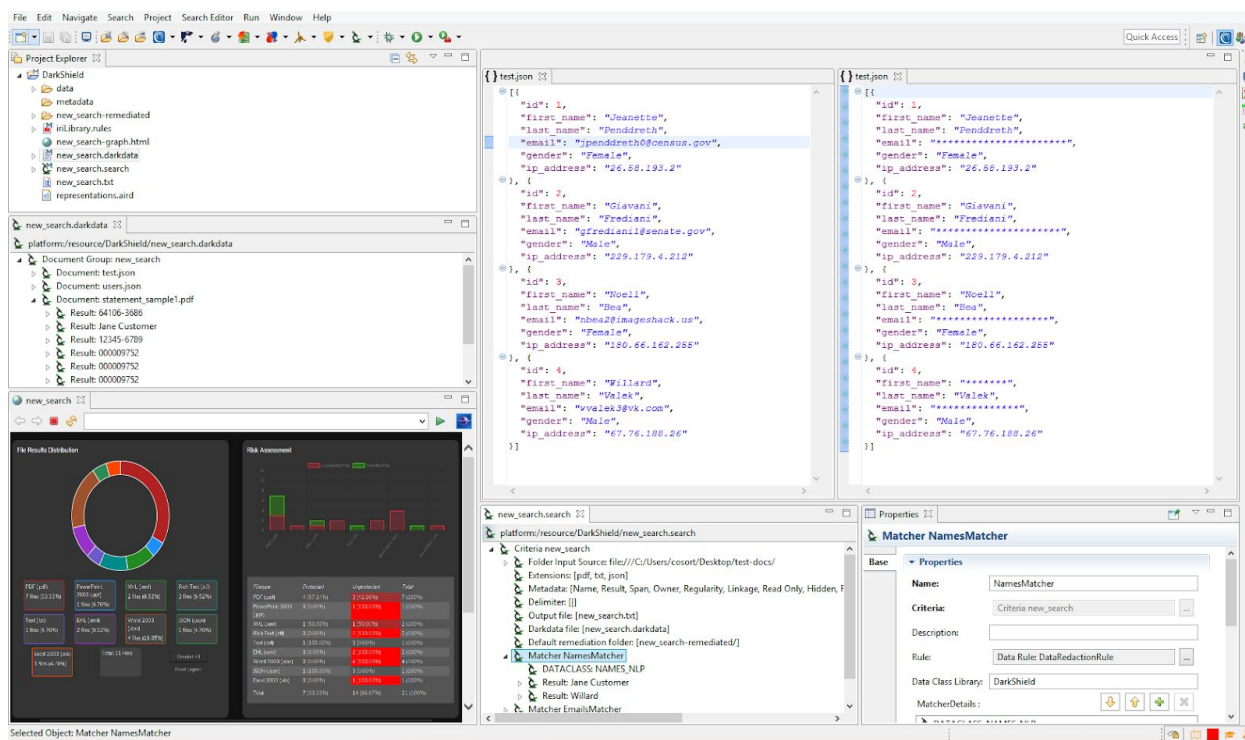
DarkShield is front-ended by default through [IRI Workbench](#), a free Graphical User Interface (GUI) for job design and management built on Eclipse™. Workbench is also where IRI FieldShield, CellShield Enterprise Edition, and other Voracity-supported, [SortCL](#)-compatible products operate.

It is also possible to invoke DarkShield jobs on the command line, or launched as an external process from any general purpose programming language. Contact IRI for help integrating DarkShield into your application environment.



For DarkShield users in particular, IRI Workbench includes:

1. The Dark Data Discovery wizard for creating searching and masking jobs for unstructured text, document, and image sources.
2. Form editors for viewing and editing data classes and DarkShield jobs
3. Click-to-run, run configuration dialog, and built- task scheduler options to launch and automate repeating DarkShield jobs that search, mask, or do both at once.
4. A Named Entity Recognition (NER) Model wizard to leverage the power of Natural Language Processing (NLP) and Machine Learning (ML) to train and use custom NER models to identify persons, organizations, locations, and other entities that cannot be easily found using a pattern or in a lookup set
5. A Facial Recognition wizard to train a model to select and blur particular faces
6. Textual and graphical views of DarkShield search and remediation results.
7. Offline and online technical documentation, learning articles and videos, and support from IRI engineers and IRI partners located in 40+ cities worldwide.



IRI Workbench and DarkShield run on Windows, Linux, and macOS platforms. DarkShield operations are expected to be performed on the same system, which need a minimum of 4GB of RAM. Contact IRI if you need to separate job design for execution.

The use of DarkShield features are outlined and explained in further detail below.

# DarkShield Workflow

These steps describe the most common (but not the only) way DarkShield is used:



Download &  
Installation

**1. Download and Install.** Obtain and open IRI Workbench, and license the back-end data masking executable(s) as directed in the installation guide. From the Workbench Help menu, install the latest DarkShield feature from the IRI tooling update site.



Data  
Classification

**2. Classify Your Data.** Define Data Classes (e.g. names, phone numbers, PINs) and Class Groups (e.g., ePHI) which require masking in the Data Classification dialog launched from the IRI Preferences menu in Workbench. Associate each class or group with the search method or methods (pattern, lookup value, NER model matches, etc.) for finding that data.



Masking  
Rules

**3. Define Data (Masking) Rules.** Open the new Data Rule wizard from the IRI Workbench File menu to choose, configure, and save one or more masking functions to an IRI Rule Library. One or more masking rules will be applied in Step 5, when you define the search and masking job and match these functions to data classes.



Data  
Sources

**4. Specify Your Sources/Targets.** Run the Dark Data Discovery wizard to identify all the folders in your file system or LAN where DarkShield-supported file formats reside. Then, specify the target directories where the masked documents will be stored. You can also select which metadata information to collect on the files holding PII, and name the extract file and delimiter for the search results. In the next page of the wizard, you will continue creating the search and masking job details that will be used to: a) scan your chosen sources and file formats for the Data Classes and Class Groups you defined; and, b) apply the masking functions in your Rule Library to them:



Rule  
Matchers

**5. Create (or Use) Search Matchers.** The mapping between Data Rules (masking functions) and Data Classes (or Data Class Groups) happens through Search Matchers. To create these matchers, browse to an existing Data Rule created in Step 3 above, or create a new one and associate it with a Data Class or Group in the Dark Data Discovery Wizard.



Job  
Execution

**6. Run the Job.** When you click Finish in the Dark Data Discovery wizard, your search specifications serialize into a .search file (used to run search-only or search and mask jobs). You can run the .search file from the project folder or Run Configuration menu. That search job creates a .darkdata file with the search results, which can be masked in a "Mask" job. A .search file can also be re-run at a later time to populate the .darkdata file with new search results -- either from new documents or existing ones that were modified since the last run. When a remediation job runs, it masks all the PII found in the searched files, and writes the masked files into the targets(s) that you specified (using the same file names).



Job  
Review

**7. Review the Results.** Each search job produces a delimited file with the PII values and file metadata that you specified in the wizard, and a .darkdata file tree view of the same. The .darkdata file can also generate an interactive dashboard view of the different file types found, as well as the number and ratio of files per category in which every search result was successfully masked. And of course, if you performed remediation (masking), you can open and review the newly masked files in their target folders. Log data can also feed SIEM tools.



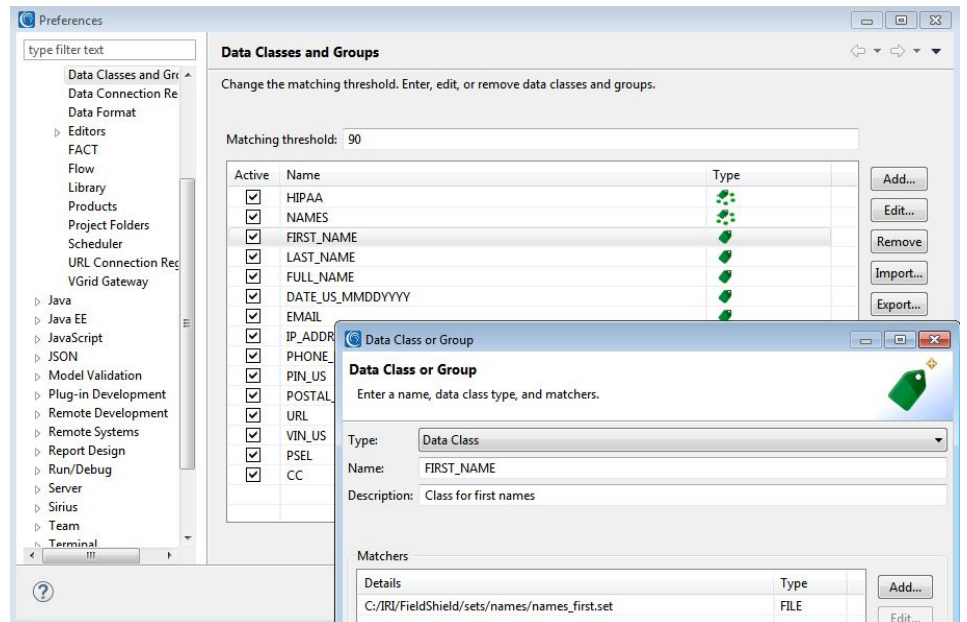
Job  
Scheduling

**8. Automate the Job.** Once you are familiar with running DarkShield and comfortable with the results it produced, you can repeat its jobs with the Workbench task scheduler or your own via CLI calls. Each time a job runs, it will perform the same searching and masking on new files in your source folders, or re-scan and mask files updated since the last search.

## Data Classification

DarkShield shares the same data classification facilities as FieldShield to define and catalog one or more items of PII.

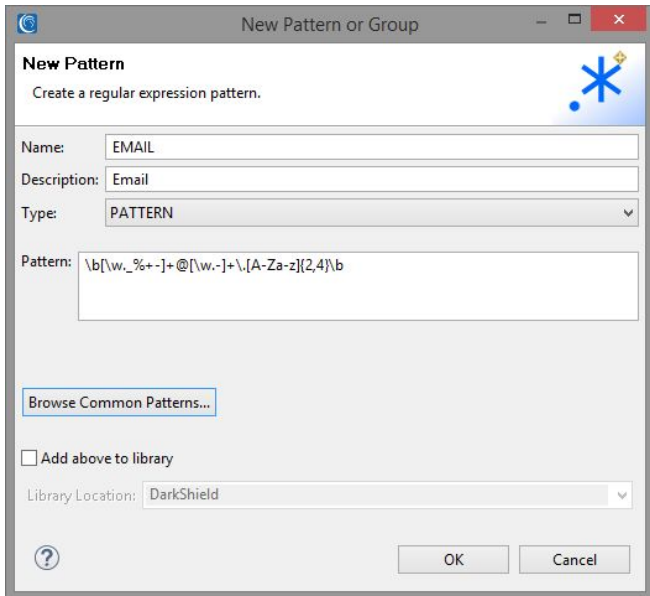
These items are classified through the use of Data Classes or Data Class Groups, which are categorized by any combination of search matchers, including:



1. Strings conforming to IRI-supplied or custom-defined Java Regular Expression (Regex) patterns, which are ideal for Personal Identifiable Numbers (PINs), email addresses and phone numbers. These Regex searches can also be computationally validated at the same time to avoid false positives.
2. Exact matches to strings in a lookup file/table (e.g., countries)
3. Named 'path' or column filters for JSON, XML, CSV, Excel
4. Named-Entity Recognition (NER), based on machine-trained Natural Language Processing (NLP) Models (e.g., names and addresses)
5. Bounding boxes to define specific, repeated regions within images to mask
6. Facial detection and recognition

Data Classes and Data Class Groups can be defined and saved or used in DarkShield, and the other two "Shield" products through global preferences in IRI Workbench.

Currently, NER models are only supported as search matchers in DarkShield jobs. Fuzzy match searches supported in FieldShield are not yet supported in DarkShield. Facial recognition and bounding boxes are only supported for image formats.



## Regular Expression Patterns

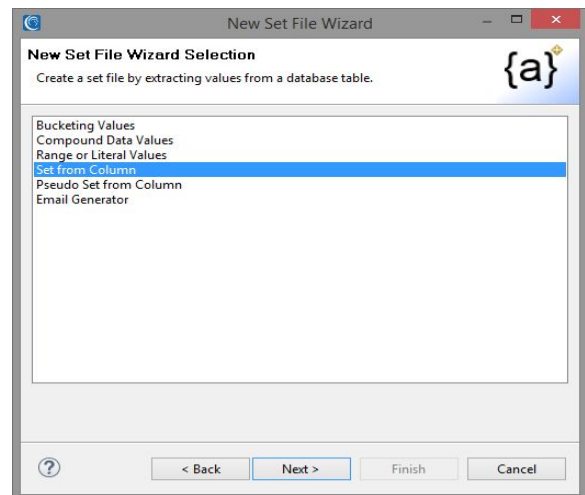
DarkShield can use any Java regular expression (RegEx) to find PII data that conforms to a well-defined format (email address, credit card number, etc.).

IRI Workbench provides many common patterns, and allows DarkShield users to create and save their own for re-use in Dark Shield data classification and other IRI searching and masking wizards, too. Patterns can be saved and shared across projects using the IRI Patterns library.

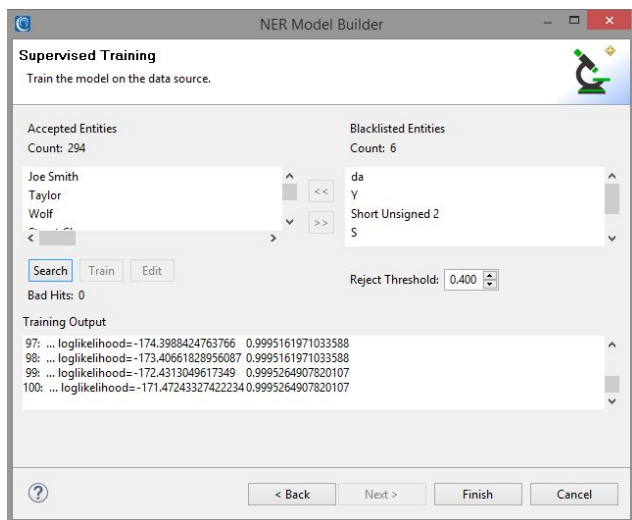
## Set File Lookups

DarkShield supports the use of set file lookups for finding names and other proper nouns through direct string matches to values in a lookup file.

DarkShield can create their own set files in IRI Workbench manually (through various Set File creation wizards), or automatically by extracting data from database columns that can be reached through a JDBC connection.



## NER Models



DarkShield supports the use of any OpenNLP Name Finder model found [here](#), or obtained afield. In cases where the pre-trained models do not provide accurate results from searches through context-specific documents, graphical user wizards in IRI Workbench help you use or train custom models for DarkShield. They can use existing annotated training data, or create it through a semi-supervised and iterative training process using your documents.

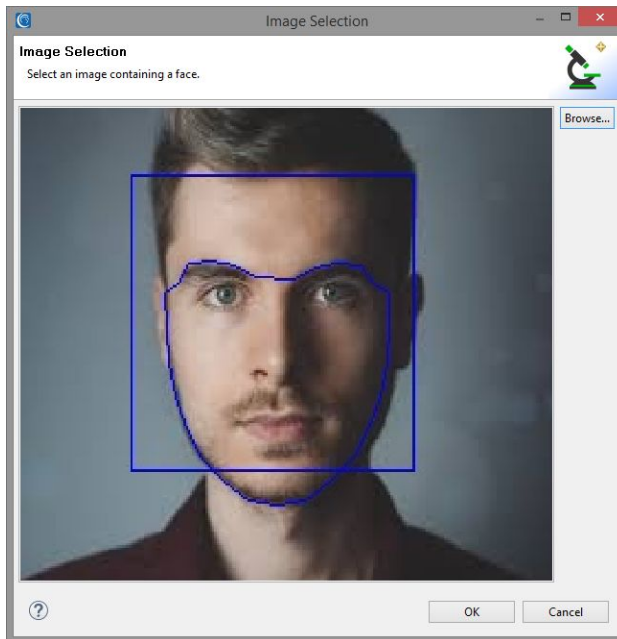
## Bounding Boxes

DarkShield supports the definition of regions within image files to be masked. This is especially useful if other PII discovery (search) methods failed, and the area in which the PII exists in one or more like files is known.

A user-friendly area drawing tool is provided in the details area of the Data Class Matcher dialog. It defines a “bounding box” around the content you want redacted in each file.



## Facial Detection & Recognition



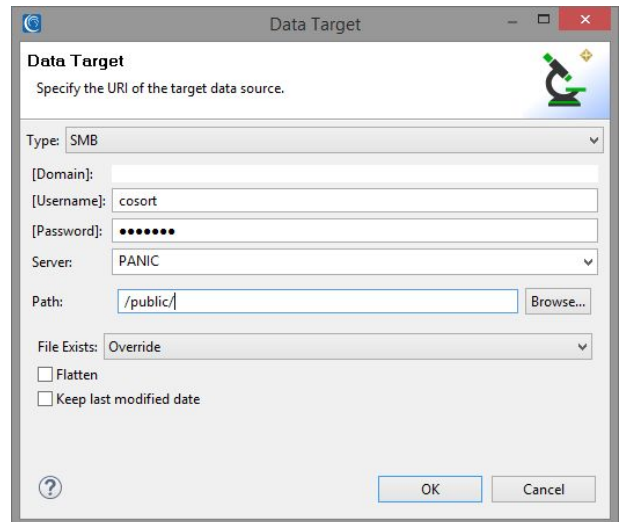
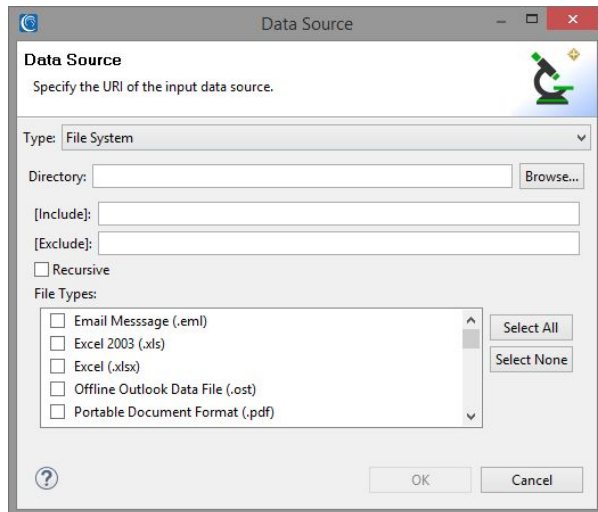
DarkShield supports the use of facial detection technology to find and blur faces located in image files.

DarkShield also supports the creation of Facial Recognition models which can be trained to find and blur only specific faces.



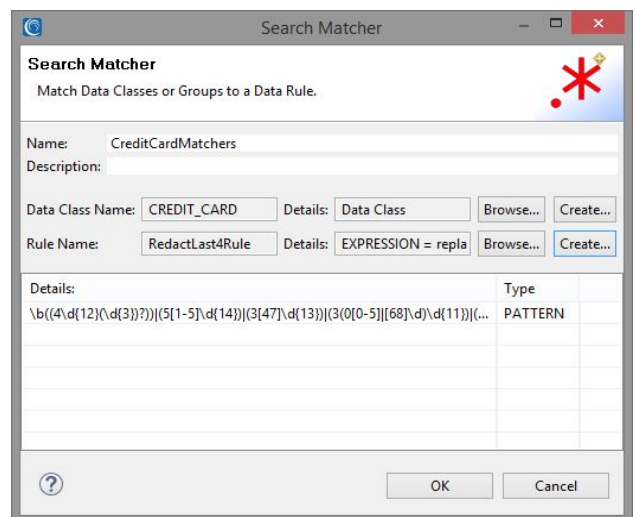
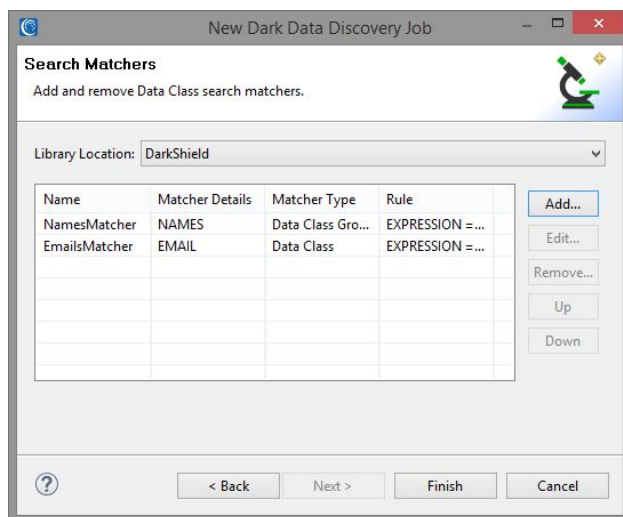
## Dark Data Discovery / PII Search

DarkShield uses the Dark Data Discovery wizard in IRI Workbench to define both PII search and masking jobs. It allows the user to specify all the file formats to be searched, and Server Message Block (SMB) share drives and folders they reside in, and the target drives and folders to store the masked files.



In addition, the wizard also allows the user to select various metadata attributes associated with each file in which PII is discovered, including its ownership, linkages, creation and last modification dates, etc. That information is included in a delimited flat-file containing all of the search results.

The search criteria associated with the Data Classes and Groups are matched to your chosen masking function by creating Search Matchers in the next page of the wizard:



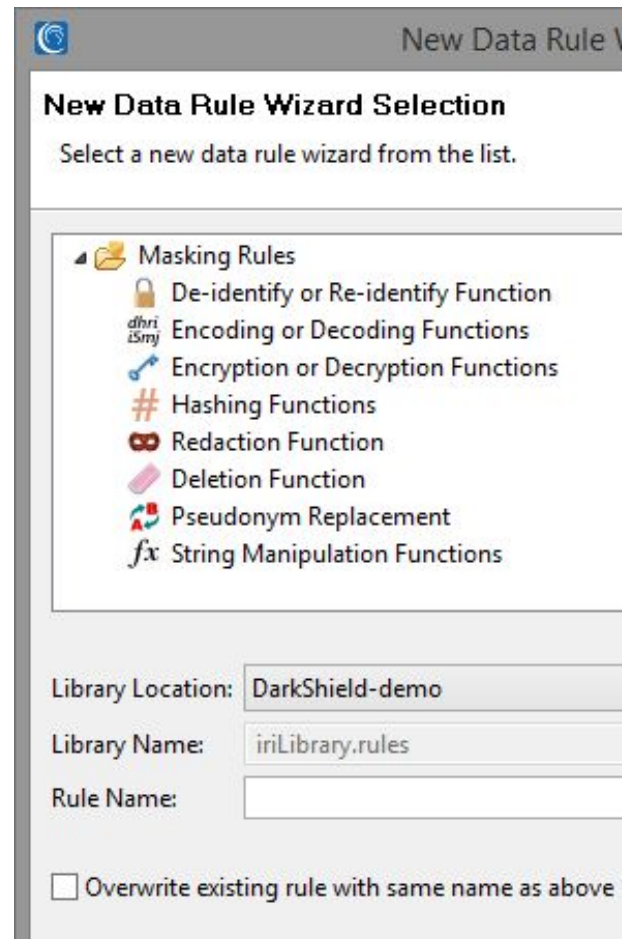
The data masking functions that can be used are described in the next section.

## Data Rules and Masking Functions

DarkShield applies masking functions by using data rules. Data rules can be created and stored for future use and modification in an IRI Rule library stored in an [IRI Workbench](#) project folder.

These Data Rules can be matched to Data Classes or pattern matchers when defining Data Rule Matchers in the Dark Data Discovery Wizard. The Data Rule Matchers are then used to consistently mask the discovered PII via:

1. multiple, NSA Suite B and FIPS-compliant encryption (and decryption) algorithms, including *format-preserving* encryption
2. SHA-1 and SHA-2 hashing
3. ASCII de-ID (bit scrambling)
4. binary encoding
5. deletion (erasure / removal)
6. redaction (full or partial string masking)
7. lookup value pseudonymization
8. byte shifting and (sub)string functions



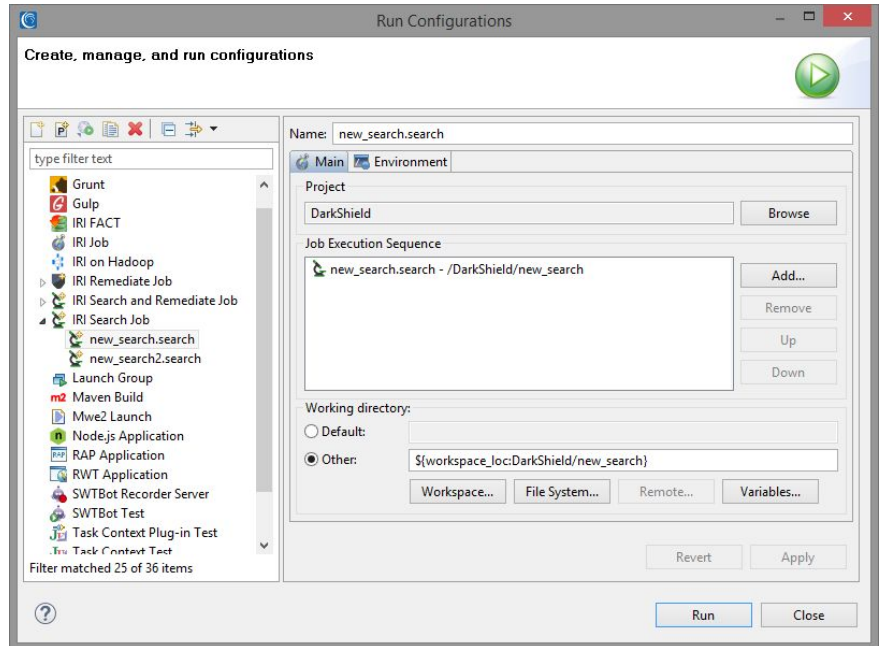
Except for pseudonymization using restore sets, DarkShield masking functions are not readily reversible. If you used encryption, encoding, or certain string functions, and deleted your unmasked source documents (so the only version left has been masked by DarkShield), contact IRI for a service-based restoration effort. You must also have the original .darkdata file for this to be possible.

DarkShield can search and mask any unstructured text and document files with all of the methods described in this booklet. However, image formats and some .pdf files only support black box (redaction) remediation or blurring (for faces) irrespective of the specified Data Rule.

[IRI CellShield Enterprise Edition \(EE\)](#) can also be used to mask data in multiple Excel spreadsheets from within a master Excel sheet and function selector dialog. CellShield EE can also mask DarkShield-discovered PII in one or more sheets at a time, and conduct further intra-cell searching and masking operations directly in Excel. CellShield provides its own subset of masking functions to mask search results specified in the fit-for-purpose .eif file created by the Dark Data Discovery wizard in IRI Workbench.

## Running DarkShield Jobs

PII searching and masking jobs are designed, managed, and run from IRI Workbench. Jobs can run in the same pass by running a “search & mask” job from the `.search` configuration file, or separately by first running a “search” job and then running a “mask” job on the `.darkdata` file generated from the search. The user can save the run configuration for both ad hoc or scheduled (repeat) executions using the built-in task scheduler in IRI Workbench.



Repeated DarkShield runs can detect changes in files that were previously searched on subsequent runs, and repeat the search.

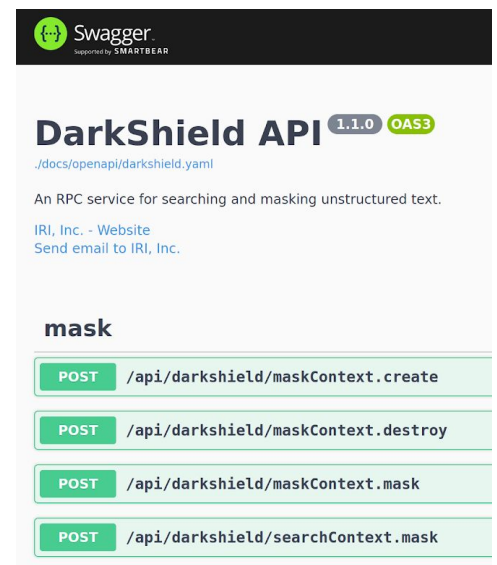
```
cmd
Command Prompt
(c) 2013 Microsoft Corporation. All rights reserved.
C:\Users\cosort>darkshield -h
Usage: darkshield [-Criteria Path [-r] [-t=Tessdata] ! [-d=Dark Data Path]]
[-h] [-h=Binary Directory] [-e=Executable]
Command Line API for the DarkShield search and remediation engine.
-h, --bin-dir=Binary Directory The path to the binary directory where the CoSort
executable is found. Defaults to "%COSORT_HOME/bin"
-e, --executable=Executable The CoSort executable to use for the remediation.
Defaults to "sortcl."
-h, --help Show this help message and exit.
-U, --version Print version information and exit.
Search Criteria Options
Criteria Path Path to the Search Criteria configuration file.
-r, --remediate Remediate the files as they are found by the search.
-t, --tessdata=Tessdata Path to the tessdata folder containing Tesseract
training files.
Dark Data Remediation Options
-d, --darkdata=Dark Data Path Path to the dark data file.
Options
C:\Users\cosort>
```

## Command Line Interface (CLI)

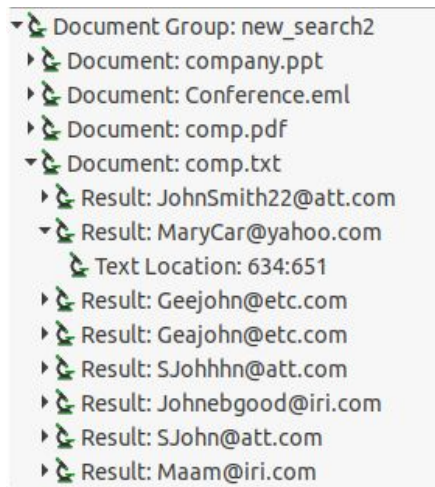
The DarkShield CLI runs file-based search and remediation jobs from outside IRI Workbench, via other programs in server environments with a Java runtime.

## Remote Procedure Call (RPC) API

The [DarkShield API](#) allows application programs and web services to call DarkShield’s powerful searching methods and masking functions for both text and file sources in a virtually unlimited range of formats and systems (subject to “glue code” customizations). Embedding this functionality allows you to bypass IRI Workbench and deploy DarkShield in more automated, and orchestrated, environments that may be distributed on-premise or in the cloud.



## Reporting and Using Results

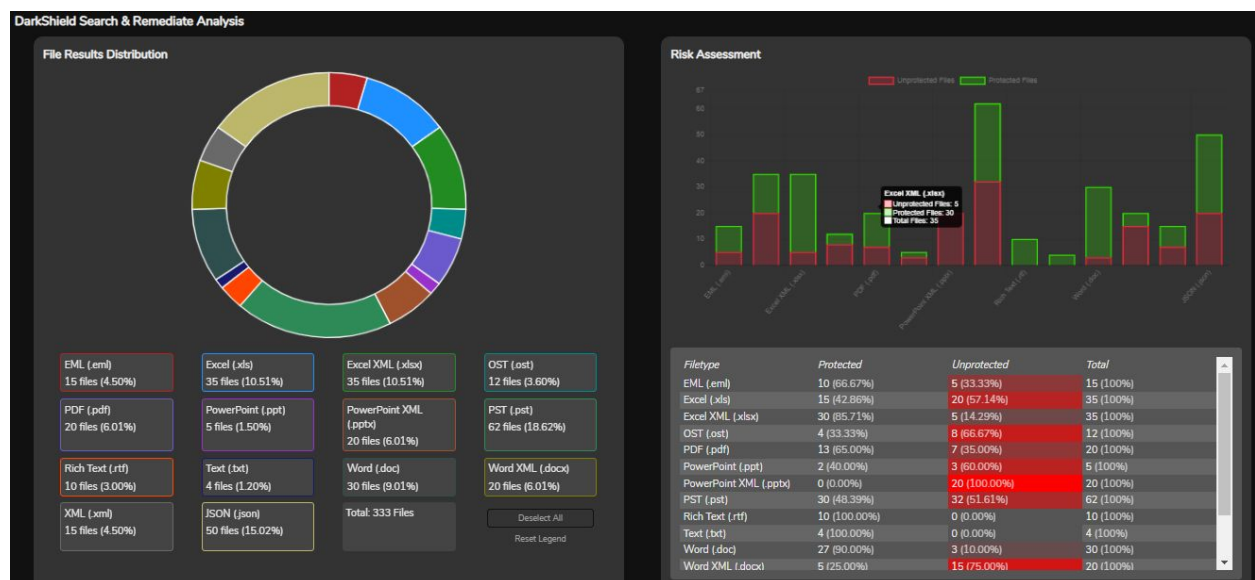


When DarkShield runs, it produces several files which can be reviewed for audit purposes, and to comply with the GDPR provision for data portability. Specifically, every search generates and updates a text file with a list of its search results and whatever user-specified metadata information on the source files was selected.

With each search, DarkShield can also create a Data Definition File ([DDF](#)), or metadata repository defining the fields you picked for the search file. In the same UI, IRI CoSort can write a [custom report](#) using that layout.

With or without a query tool like the CoSort [SortCL](#) program, you have extracted the PII values matching your search criteria, so you can delete and/or provide them to auditors. For GDPR compliance, you can also provide the results of individual name searches to those requesting “data portability” and “the right to be forgotten.” You will be able to show them what data about them was found, and what data was deleted.

You can determine what data was deleted through the .darkdata file, which is also produced after a search, or a search and mask operation. The .darkdata file contains a list of documents that were searched, along with the search results found under each document. You can send this log to a SIEM tool like Splunk ES, or directly open a graph in IRI Workbench showing the sensitive data found, and what was or was not masked:

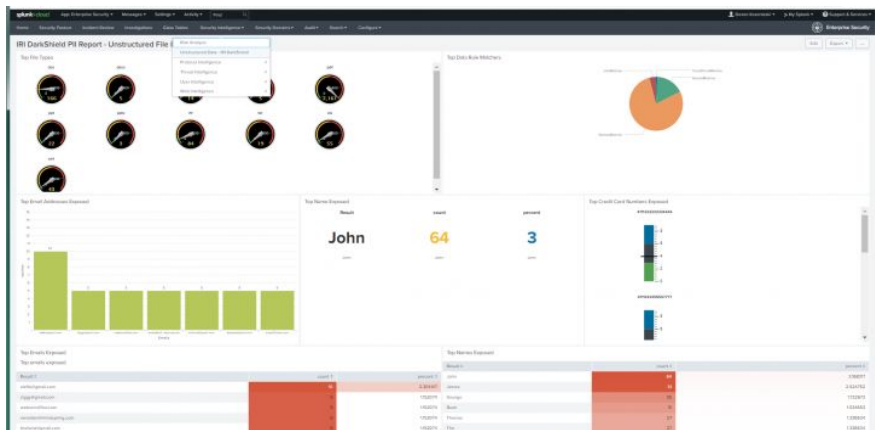


*DarkShield graphical report generated in IRI Workbench showing the distribution of file types containing the search results, and a risk assessment of how many files are protected in each format category.*

## SIEM Tool Integration

Security Information Event Management (SIEM) tools like Splunk Enterprise Security (ES) and modern analytic platforms like Datadog designed to create insights and enable actions from machine log data. As such, they can be used to categorize, graphically reveal, and report on security incidents from data in log files.

DarkShield produces a high volume and quality of log file data from its PII search and mask operations. The flat-file logs produced by [DarkShield can feed Splunk ES](#) directly. This supports insight into PII-related vulnerabilities in the files searched on the network, as well as those in which DarkShield has fully masked the PII it found.



For example, you can design graphical widgets that use the results of discrete log queries in Splunk, and arrange them inside a dashboard accessed from a URL. Custom views can thus reveal fine details about the PII DarkShield finds.

With that data indexed in Splunk, it is also possible to leverage the Adaptive Response Framework in the ES version to [send alerts](#) or run Phantom Playbooks [based on conditions](#) detected in DarkShield logs. For example, an email can be sent when a certain number of files with unmasked PII was recorded, thus telling DarkShield or its user to run or re-run a data masking job against the current search results.

As DarkShield searching and masking jobs are generated, they can also be [sent to Datadog](#) or [forwarded to Splunk automatically](#). That updates the DarkShield log data indexed in Splunk, and can thus trigger new response actions -- like a dashboard refresh or new alert email.

## File Formats & Databases Supported

DarkShield v4 is able to find and mask PII in the data sources listed here, whether they are in local or LAN-connected file systems (including DropBox) or Amazon S3 buckets:

Text	Documents	Images	Databases
.asc	.doc/x	.bmp	RDBs via JDBC
.html & .eml	.ppt/x	.gif	MongoDB
.hl7 & .x12	.xls/x	.jpg/x/2	CassandraDB
.json & .xml	.pdf	.png	Elasticsearch
.txt	.rtf ( <i>search only</i> )	.tiff	HIVE via API

DarkShield v4 can optionally detect faces in image files -- and recognize particular faces through computer-assisted training -- in order to blur them. Contact IRI if you need support for DICOM medical images, Outlook archives, CAD drawings or other formats.

## Data Silos in Development

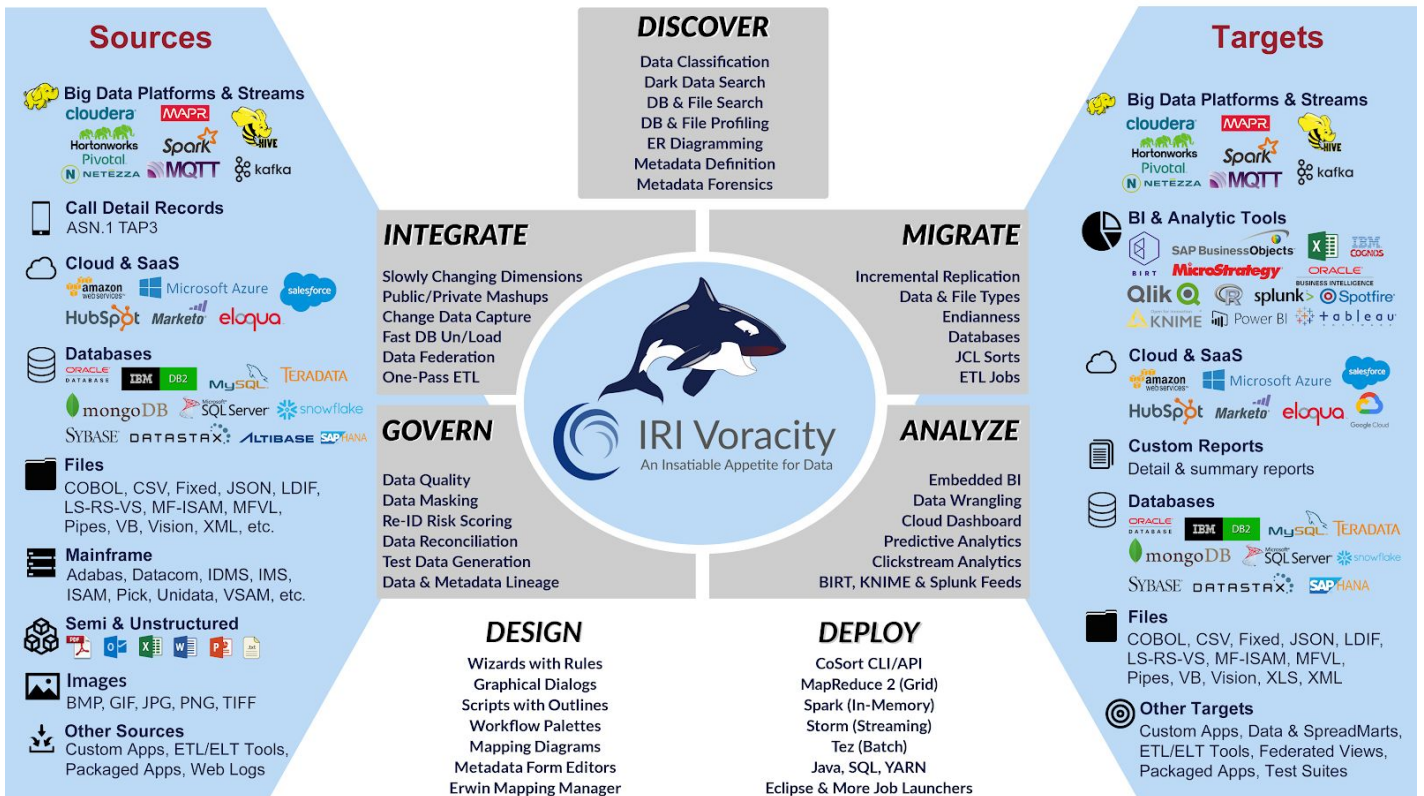
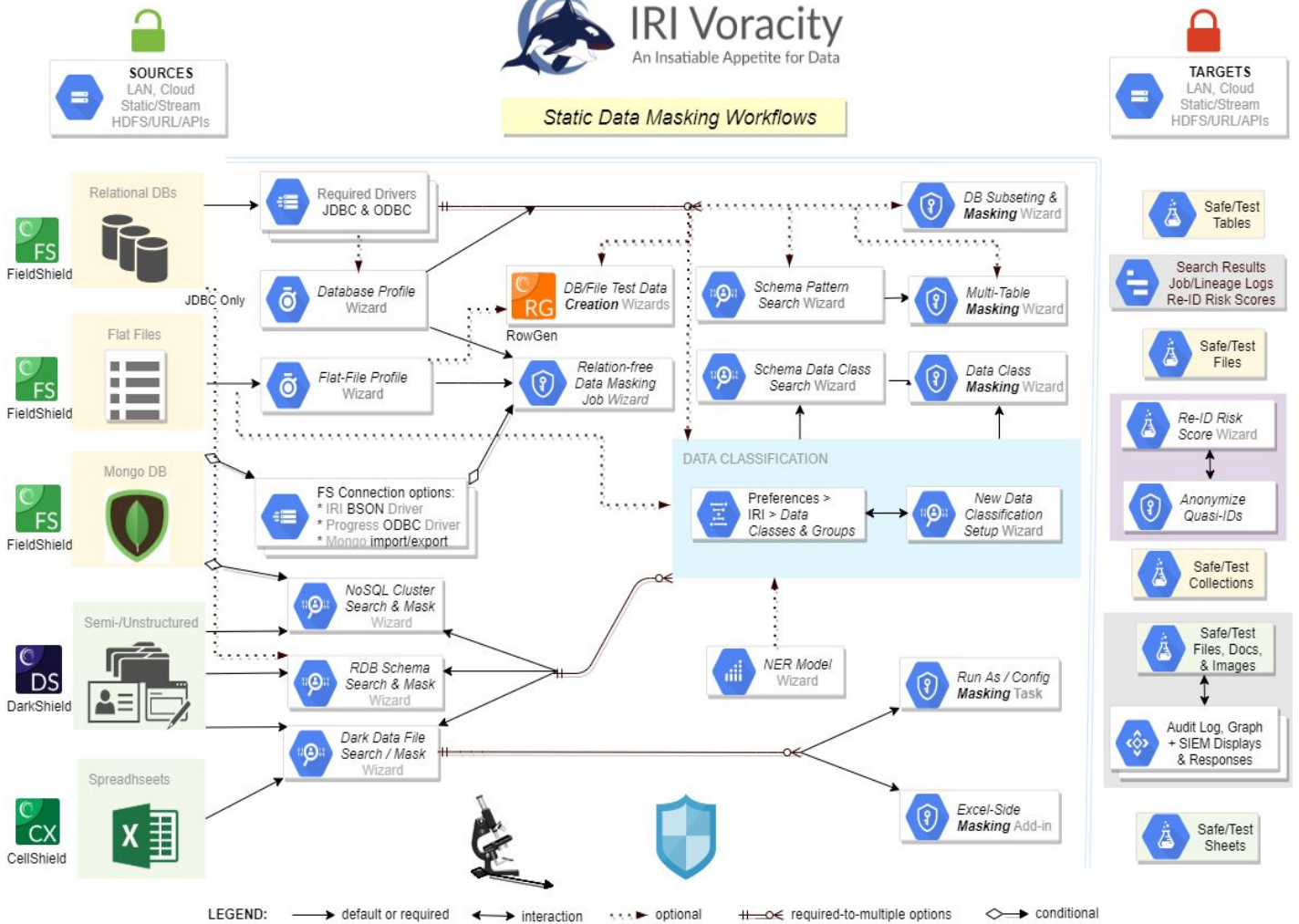
LAN, Related	Amazon	More Clouds/Apps	Additional Sources
ActiveMQ	CloudWatch	Box & Salesforce	Apache CXF & Ignite
FTP/HTTP/MINA	Dynamo	Facebook & LinkedIn	Couch DB
Google Drive	Redshift	Google Apps	HBASE & HDFS
Mainframes	SES & SNS	Google BigQuery	JDBC & JPA
Sharepoint	SQS & SWF	jclouds	Kafka & MQTT

*If your file format or silo is not on the list above, please contact [darkshield@iri.com](mailto:darkshield@iri.com) to ask if it has been added since the publication of this booklet, or when it could be added.*

## Compatible Platforms and Applications

DarkShield runs on Windows, Linux, and macOS platforms. It uses the same IRI Workbench IDE, data classes, and masking engines as:

- IRI FieldShield - DB and flat-file masking
- IRI CellShield EE - Excel spreadsheet masking
- IRI CoSort - Data transformation and reporting
- IRI Voracity - Big data management, ETL, etc.





**INNOVATIVE ROUTINES INTERNATIONAL (IRI), INC.**

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