IRI DarkShield
Unstructured Data Search & Security

Product Overview

Technical Summary, Samples, and Specifications

IRI
Total Data Management

Executive Summary

In an information technology era when both big data opportunities and privacy laws exist and converge, there is a pressing need to discover, work with, and protect data hidden in unstructured sources. Data in these silos 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, emails and other free text, plus documents, images, and audio/video files. Like transactional data in structured sources, the information contained in semi- and 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 CoSort-driven 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

IRI DarkShield
Unstructured Data Search & Security

database
TRENDS AND APPLICATIONS
TREND-SETTING
PRODUCTS

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 can use any combination of regular expressions, value lookups, path filters, and Named Entity Recognition (NER) models to search for PII floating in: relational and NoSQL database collections; Microsoft Office documents; PDF, JSON, XML, and other EDI and plain-text files like logs and emails;, plus, most image file formats through Optical Character Recognition (OCR). DarkShield also supports bounding box areas and facial recognition to positionally redact PII in images. Support for compressed, A/V, and other proprietary application formats may be added in the future.

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 their associated locational data. Supported masking functions include: redaction, encryption, pseudonymization, hashing, encoding, bit and string manipulation, random noise (blurring), scrambling, randomization, and value deletion.

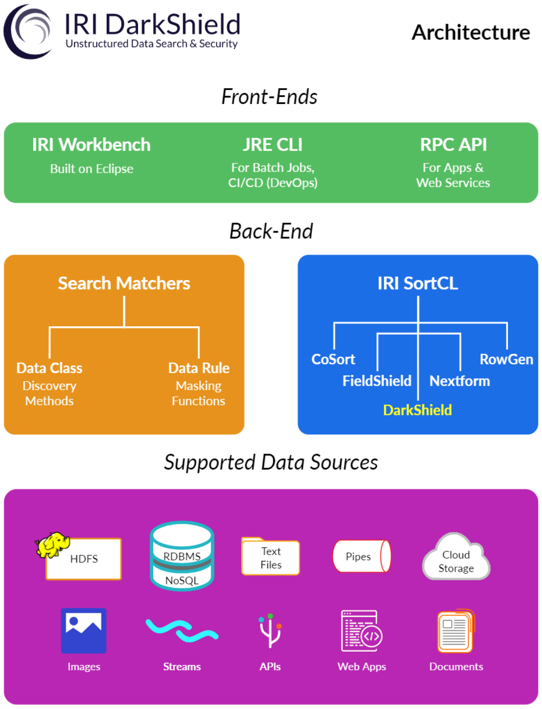
DarkShield-process metadata is serialized in JSON (API) or XMI (Eclipse GUI) for easy modification and repeat use, and can be shared in cloud repositories like Git. Search and masking results are in a flat-file log that can be audited in several ways, including: ad hoc, via built-in interactive dashboard, 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.

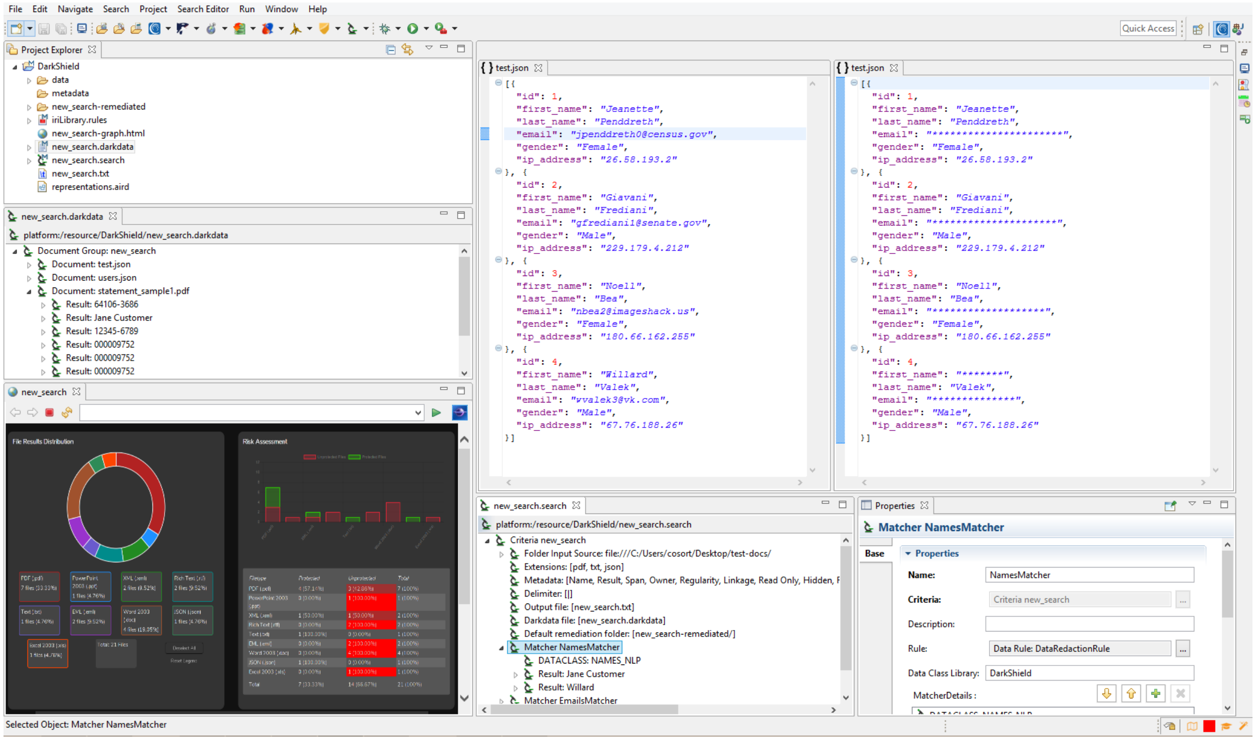
DarkShield jobs are configured through either IRI Workbench, a free Graphical User Interface (GUI) for job design and management built on Eclipse™, or the DarkShield RPC API for text or files. Workbench is also where IRI FieldShield, CellShield Enterprise Edition, and other SortCL-compatible tools operate.

The DarkShield API also makes it possible to integrate with document management and other software systems, and leverage external pre-processing tools, load balancing, and user authentication.



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.



Project Explorer:

- DarkShield
  - data
  - metadata
  - new_search-remediated
  - iriLibrary.rules
  - new_search-graph.html
  - new_search.darkdata
  - new_search.search
  - new_search.txt
  - representations.aird

new_search.darkdata

platform:/resource/DarkShield/new_search.darkdata

- Document Group: new_search
  - Document: test.json
  - Document: users.json
  - Document: statement_sample1.pdf
    - Result: 64106-3686
    - Result: Jane Customer
    - Result: 12345-6789
    - Result: 000009752
    - Result: 000009752
    - Result: 000009752

test.json

```
[{
  "id": 1,
  "first_name": "Jeanette",
  "last_name": "Penddreth",
  "email": "jpenddreth0@census.gov",
  "gender": "Female",
  "ip_address": "26.58.193.2"
}, {
  "id": 2,
  "first_name": "Giavani",
  "last_name": "Frediani",
  "email": "gfrediani1@senate.gov",
  "gender": "Male",
  "ip_address": "229.179.4.212"
}, {
  "id": 3,
  "first_name": "Noell",
  "last_name": "Bea",
  "email": "nbea2@imageshack.us",
  "gender": "Female",
  "ip_address": "180.66.162.255"
}, {
  "id": 4,
  "first_name": "Willard",
  "last_name": "Valek",
  "email": "wvalek3@vk.com",
  "gender": "Male",
  "ip_address": "67.76.188.26"
}]
```

test.json

```
[{
  "id": 1,
  "first_name": "Jeanette",
  "last_name": "Penddreth",
  "email": "**********************",
  "gender": "Female",
  "ip_address": "26.58.193.2"
}, {
  "id": 2,
  "first_name": "Giavani",
  "last_name": "Frediani",
  "email": "*********************",
  "gender": "Male",
  "ip_address": "229.179.4.212"
}, {
  "id": 3,
  "first_name": "Noell",
  "last_name": "Bea",
  "email": "*******************",
  "gender": "Female",
  "ip_address": "180.66.162.255"
}, {
  "id": 4,
  "first_name": "*******",
  "last_name": "Valek",
  "email": "**************",
  "gender": "Male",
  "ip_address": "67.76.188.26"
}]
```

new_search.search

platform:/resource/DarkShield/new_search.search

- Criteria new_search
  - Folder Input Source: file:///C:/Users/cosort/Desktop/test-docs/
  - Extensions: [pdf, txt, json]
  - Metadata: [Name, Result, Span, Owner, Regularity, Linkage, Read Only, Hidden, F
  - Delimiter: [|]
  - Output file: [new_search.txt]
  - Darkdata file: [new_search.darkdata]
  - Default remediation folder: [new_search-remediated/]
  - Matcher NamesMatcher
    - DATACLASS: NAMES_NLP
    - Result: Jane Customer
    - Result: Willard
  - Matcher EmailsMatcher

Properties

Matcher NamesMatcher

Properties

Name: NamesMatcher

Criteria: Criteria new_search

Description:

Rule: Data Rule: DataRedactionRule

Data Class Library: DarkShield

MatcherDetails :

Selected Object: Matcher NamesMatcher

IRI Workbench and DarkShield run on Windows, Linux, and macOS platforms, on physical or virtual nodes, as well as in containers, which you host on-premise or in the cloud. DarkShield operations should be staged on a system with a minimum of 4GB of RAM, but preferably 64GB.

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 or the DarkShield API, and license the back-end data masking executable(s) per either IRI installation guide. From the Workbench Help menu, install the latest DarkShield feature from the IRI tooling update site.

Data Classification

2. ***Classify Your Data (GUI only)***. 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.)

.

Masking Rules

3. ***Define Data (Masking) Rules***. Set up a DarkShield Mask Context in the API, or the new Data Rule wizard from the IRI Workbench File menu, to select, configure, and save one or more masking functions to an IRI Dark Data Rule Library. One or more masks 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****.* Write a DarkShield API calling program (a/k/a "glue code"), or run the Dark Data Discovery wizard in Workbench, to identify the DB silos or folders in your file system, LAN or cloud store where DarkShield-supported sources reside. Then, specify the target folder or schema where masked data (in the same format) will go.

Rule Matchers

5. ***Create (or Use) Search Matchers***. The mapping between Data Rules (masking functions) and Data Classes/Groups (in Workbench) 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. If you are using the DarkShield API, see how Search and Mask Contexts build and co-relate.

Job Execution

6. ***Run the Job****.* When you click Finish in the Dark Data Discovery wizard or run your API calling program, your search specifications serialize into a *.search* (configuration) file in Workbench, or in a JSON annotation file in the API. You can then run the .search file from the project folder or Run Configuration menu, to create a *.darkdata* file in Workbench containing the search results and masking rules ready to run in a combined search/mask job or separate mask job later.. The .search file can also be re-run at a later time to populate the .darkdata file with new search results. Masking jobs affect all associated PII found in the search, and write masked data to new target folders or DB collections with the same formats and names. Similar process options are available through the DarkShield API.

Job Review

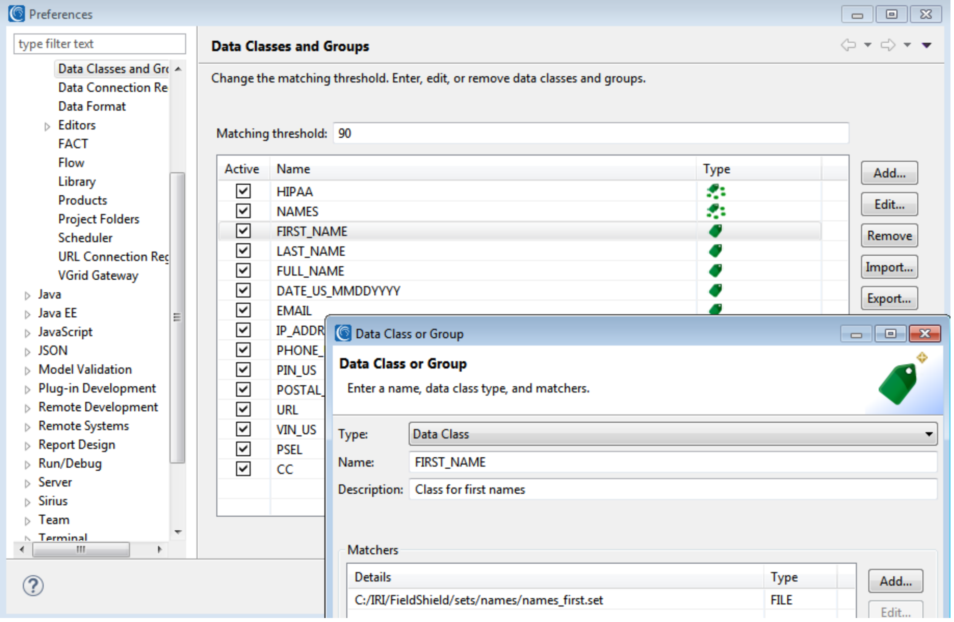
7. ***Review the Results***. Each search job in Workbench produces a delimited file with the PII values and file metadata specified in the wizard, plus a .darkdata file tree view of the same and an optional interactive dashboard view of the different file types found and the number and ratio of files per category in which every search result was successfully masked. If you performed masking, you can also open and review the newly masked data in their target folders or tables. Log data from Workbench or API runs 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 data in your sources, or re-scan and mask data 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.



Data Classes and Groups

Change the matching threshold. Enter, edit, or remove data classes and groups.

Matching threshold: 90

Active	Name	Type
☑	HIPAA	
☑	NAMES	
☑	FIRST_NAME	
☑	LAST_NAME	
☑	FULL_NAME	
☑	DATE_US_MMDDYYYY	
☑	EMAIL	
☑	IP_ADDR	
☑	PHONE_	
☑	PIN_US	
☑	POSTAL_	
☑	URL	
☑	VIN_US	
☑	PSEL	
☑	CC	

Data Class or Group

Enter a name, data class type, and matchers.

Type: Data Class

Name: FIRST_NAME

Description: Class for first names

Matchers

Details	Type
C:/IRI/FieldShield/sets/names/names_first.set	FILE

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). The DarkShield API also supports non-exact, or ‘fuzzy' matches to those values as well.

3. Named ‘path’ or column filters for JSON, XML, CSV, Excel

4. Named-Entity Recognition (NER), based on machine-trained Natural Language Processing (NLP) or TensorFlow/PyTorch models (e.g., for words like names)

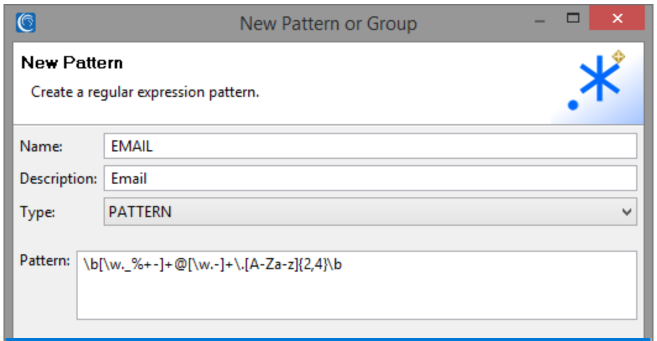
5. Bounding boxes to define specific, repeated regions within images to mask

6. Facial detection and recognition (upon request only)

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 supported in the DarkShield API.

Regular Expression Patterns


New Pattern or Group

New Pattern

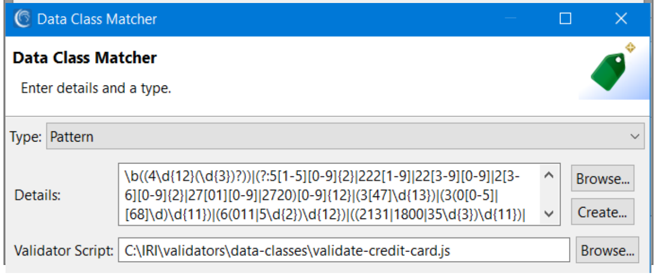
Create a regular expression pattern.

Name: EMAIL

Description: Email

Type: PATTERN

Pattern: `\b[\w._%+-]+@[\w.-]+\.[A-Za-z]{2,4}\b`


Data Class Matcher

Data Class Matcher

Enter details and a type.

Type: Pattern

Details:

```
\b((4\d{12}(\d{3})?))|(?:5[1-5][0-9]{2}|222[1-9]|22[3-9][0-9]|2[3-
6][0-9]{2}|27[01][0-9]|2720)[0-9]{12}|(3[47]\d{13})|(3(0[0-5]|
[68]\d)\d{11})|(6(011|5\d{2})\d{12})|((2131|1800|35\d{3})\d{11})|
```

Validator Script: C:\IRI\validators\data-classes\validate-credit-card.js

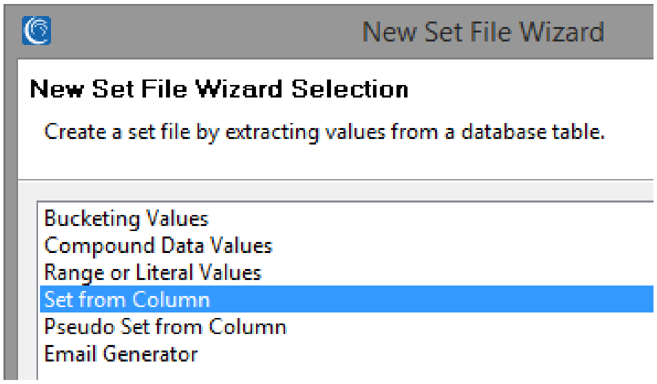
DarkShield can use any Java regular expression (RegEx) to find PII data that conforms to a well-defined format (email address, credit card number, etc.), along with further validation logic in Javascript.

IRI Workbench ships with many common patterns, and allows DarkShield users to create and save their own patterns for re-use in other IRI data classification, searching and masking wizards and projects, too.

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.

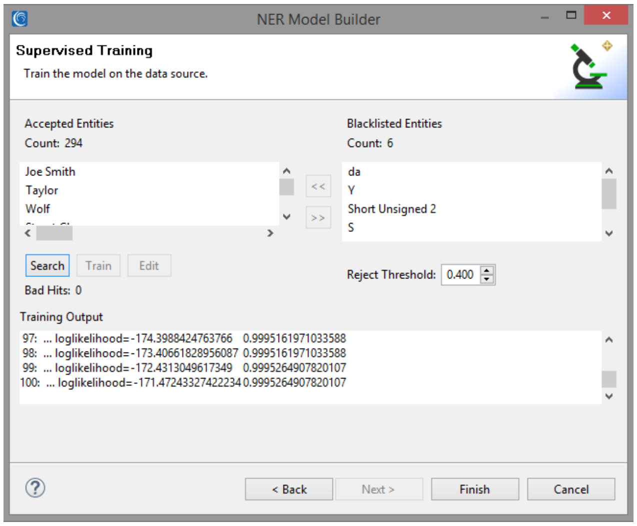
DarkShield can create their own set files in IRI Workbench manually (through various Set File creation wizards), or automatically by extracting data from database columns that can be reached through a JDBC connection.


New Set File Wizard

New Set File Wizard Selection

Create a set file by extracting values from a database table.

NER Models


NER Model Builder

Supervised Training

Train the model on the data source.

Accepted Entities

Count: 294

Joe Smith
Taylor
Wolf

Blacklisted Entities

Count: 6

da
Y
Short Unsigned 2
S

Bad Hits: 0

Reject Threshold: 0.400

Training Output

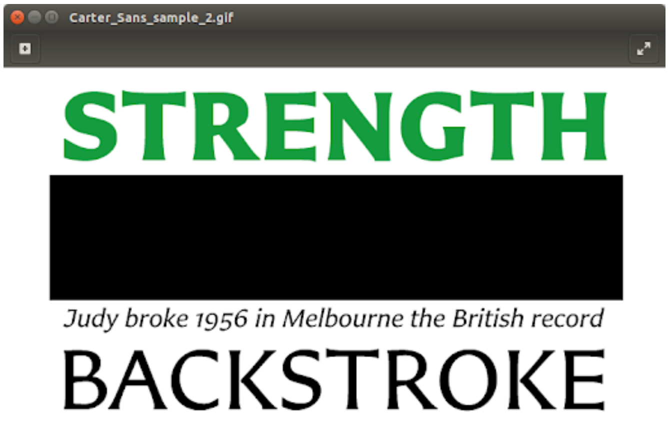
```
97:  ... loglikelihood=-174.3988424763766   0.9995161971033588
98:  ... loglikelihood=-173.40661828956087 0.9995161971033588
99:  ... loglikelihood=-172.4313049617349   0.9995264907820107
100:  ... loglikelihood=-171.47243327422234 0.9995264907820107
```

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.



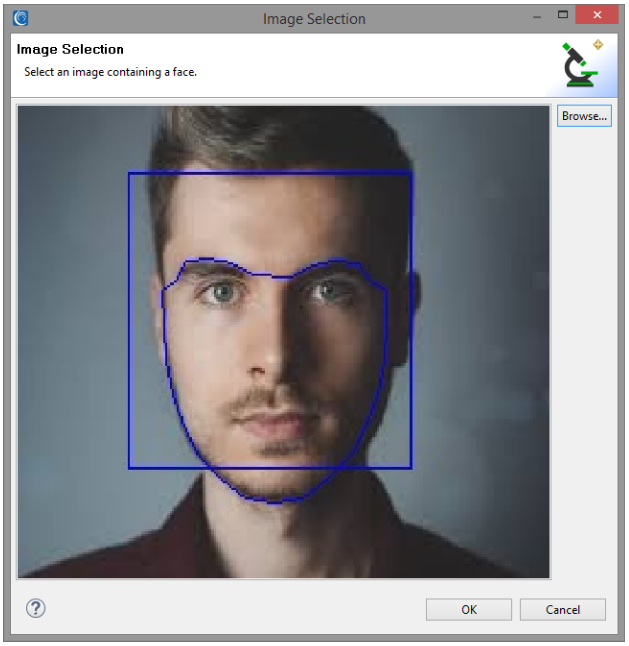
STRENGTH

[redacted]

Judy broke 1956 in Melbourne the British record

BACKSTROKE

Facial Detection & Recognition (Optional)



Select an image containing a face.

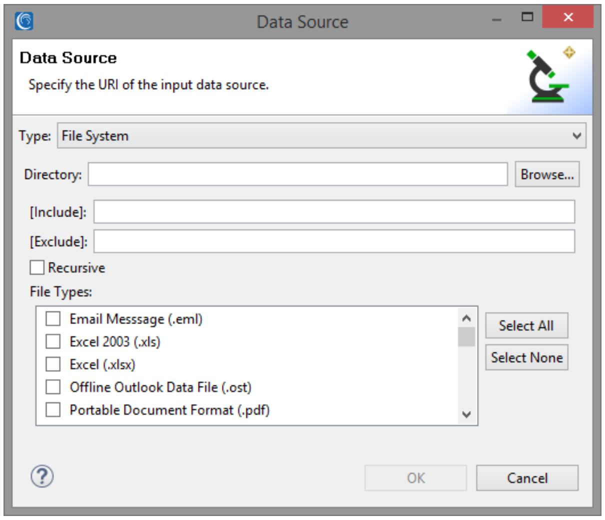
DarkShield supports the use of facial detection technology to find and blur faces located in image files.

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

Note that in DarkShield Version 4, this feature is not supplied by default, and is more involved technically. It thus may be furnished on request through Workbench or the API based on the business case.

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.



Data Source

Specify the URI of the input data source.

Type: File System

Directory:

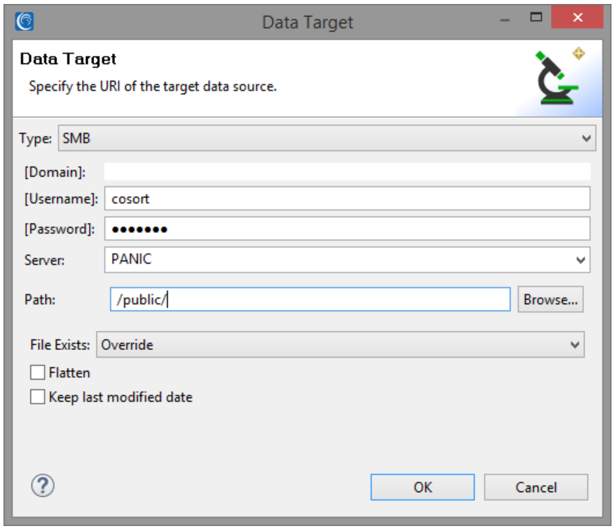
[Include]:

[Exclude]:

□ Recursive

File Types:

- □ Email Messsage (.eml)
- □ Excel 2003 (.xls)
- □ Excel (.xlsx)
- □ Offline Outlook Data File (.ost)
- □ Portable Document Format (.pdf)



Data Target

Specify the URI of the target data source.

Type: SMB

[Domain]:

[Username]: cosort

[Password]: •••••••

Server: PANIC

Path: /public/

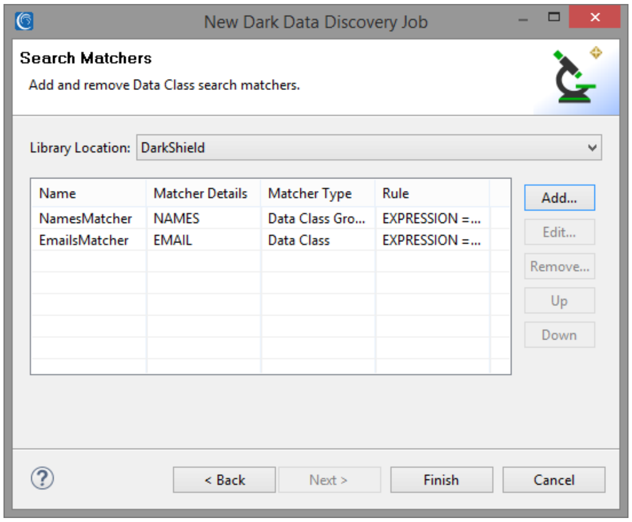
File Exists: Override

□ Flatten

□ Keep last modified date

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:

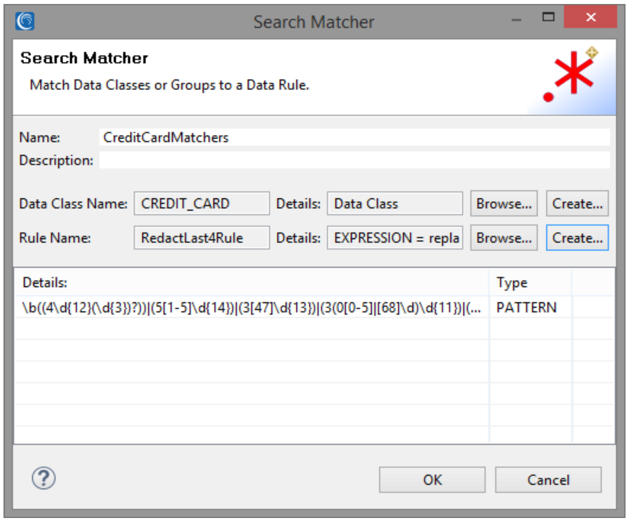


Search Matchers

Add and remove Data Class search matchers.

Library Location: DarkShield

Name	Matcher Details	Matcher Type	Rule
NamesMatcher	NAMES	Data Class Gro...	EXPRESSION =...
EmailsMatcher	EMAIL	Data Class	EXPRESSION =...



Search Matcher

Match Data Classes or Groups to a Data Rule.

Name: CreditCardMatchers

Description:

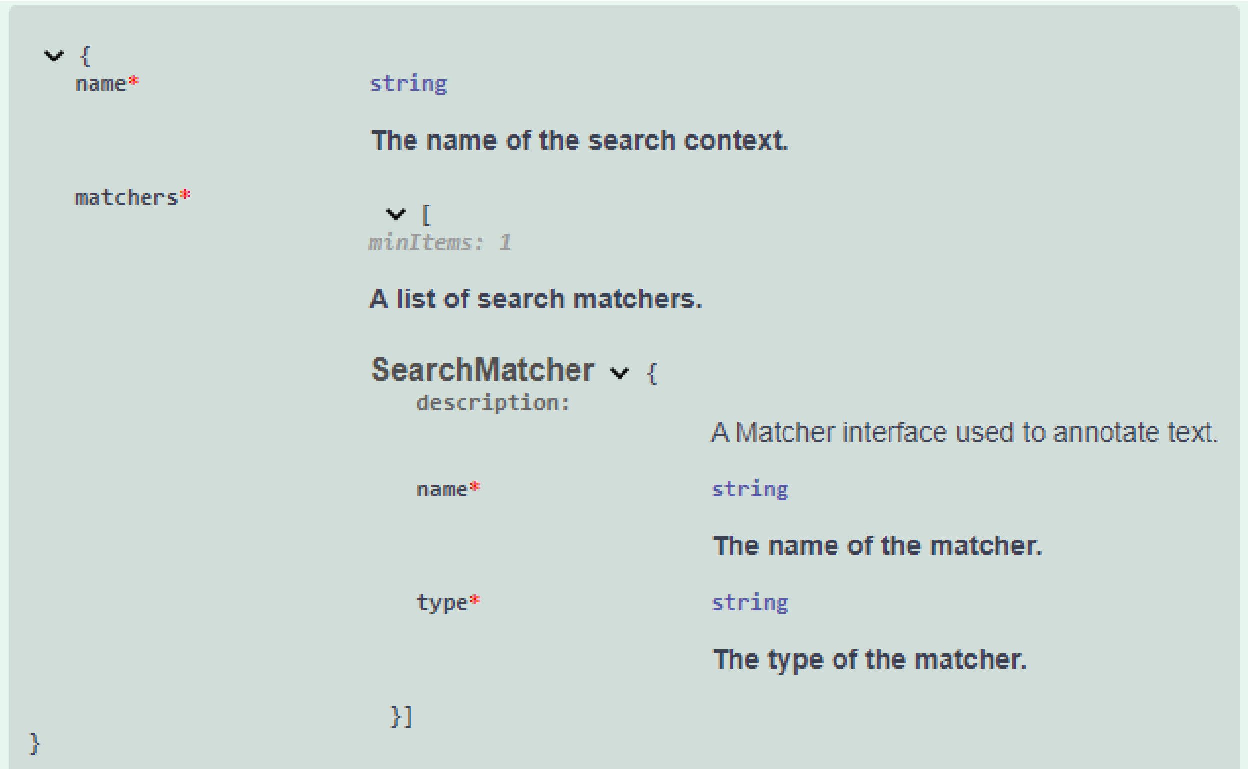
Data Class Name: CREDIT_CARD Details: Data Class

Rule Name: RedactLast4Rule Details: EXPRESSION = repla

Details:	Type
\b((4\d{12}(\d{3})?))|(5[1-5]\d{14})|(3[47]\d{13})|(3(0[0-5]|[68]\d)\d{11})|(...	PATTERN

Details of search matchers such as a regular expression pattern, validator script, set file location, NER model location, or bounding box region can be copied from here to use with the API.

With the DarkShield API, search matchers are set up as a part of a “search context.” Multiple search matchers can be grouped into a search context. This is done by sending a request to the *api/darkshield/searchContext.create* endpoint which sets a name for the context, and describes the search matchers to use.



```
{
  name*                 string

                        The name of the search context.

  matchers*             [
                        minItems: 1

                        A list of search matchers.

                        SearchMatcher {
                            description:
                                              A Matcher interface used to annotate text.

                            name*             string

                                              The name of the matcher.

                            type*             string

                                              The type of the matcher.

                         }]
}
```

For file-specific matchers such as JSON path matchers, the definition of a file-specific matcher is set up in the array of matchers as a part of a JSON request to the *api/darkshield/files/fileSearchContext.create* endpoint.

The creation and destruction of search contexts can either be delegated as part of a calling program, or handled outside the scope of the program through methods such as the Swagger UI, cURL requests, or Postman. If handled as part of a calling program, the typical flow is to create contexts at the beginning of the program, send any text or files to the API, and destroy the contexts when the program is about to complete.

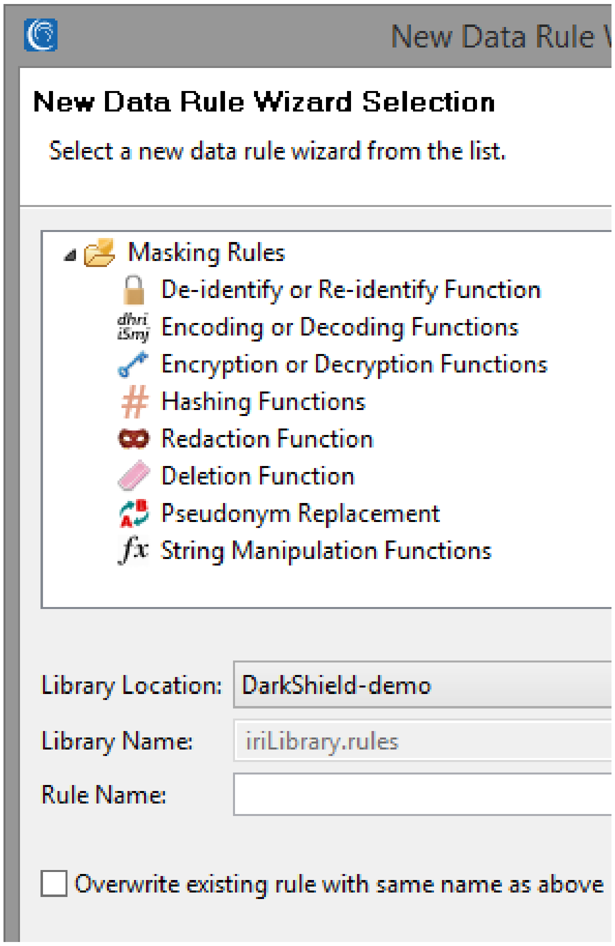
The data masking functions that can be used are described in the next section.

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


New Data Rule
Masking Rules
De-identify or Re-identify Function
Encoding or Decoding Functions
Encryption or Decryption Functions
Hashing Functions
Redaction Function
Deletion Function
Pseudonym Replacement
String Manipulation Functions

New Data Rule Wizard Selection

Select a new data rule wizard from the list.

Library Location: DarkShield-demo

Library Name: iriLibrary.rules

Rule Name:

☐ Overwrite existing rule with same name as above

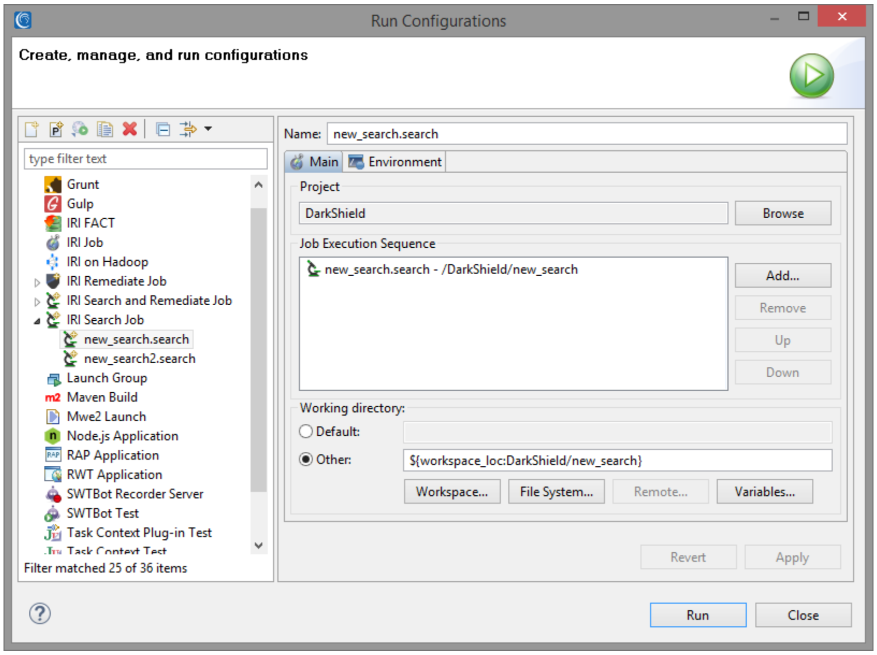
Except for pseudonymization using restore sets, DarkShield masking functions are not readily reversible. If you used encryption, encoding, or certain string functions, and deleted your unmasked source documents (so the only version left has been masked by DarkShield), contact IRI for a service-based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

For the DarkShield API, mask contexts are set up to associate search matchers with masking functions by sending a request to create a “mask context” through the *api/darkshield/maskContext.create* endpoint. For files, a file mask context must also be created through the *api/darkshield/files/fileMaskContext.create* endpoint, which allows for the specification of file-type-specific masking options.

Search and mask contexts that have been created are referenced by the name of each context when searching and masking, respectively. Examples of how to set up common search matchers and masking rules with the DarkShield API are available from the DarkShield API demos GitHub repository.

Running DarkShield Jobs

PII searching and masking jobs can be 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 configuration can be saved for ad hoc or scheduled executions.



Run Configurations

Create, manage, and run configurations

Name: new_search.search

Project: DarkShield

Job Execution Sequence: new_search.search - /DarkShield/new_search

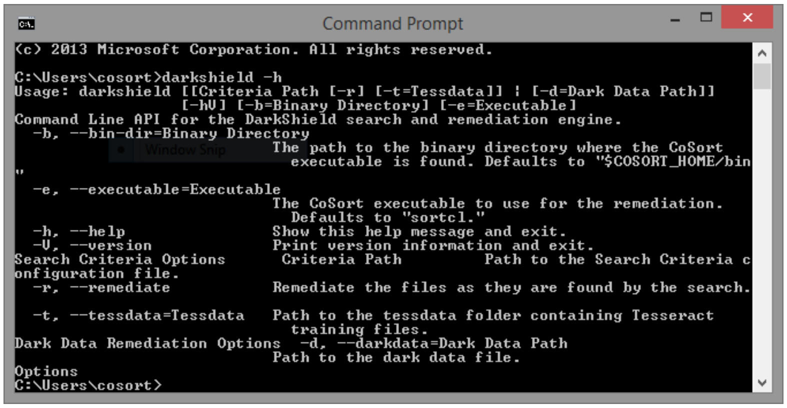
Working directory:
○ Default:
● Other: ${workspace_loc:DarkShield/new_search}

Filter matched 25 of 36 items

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

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.

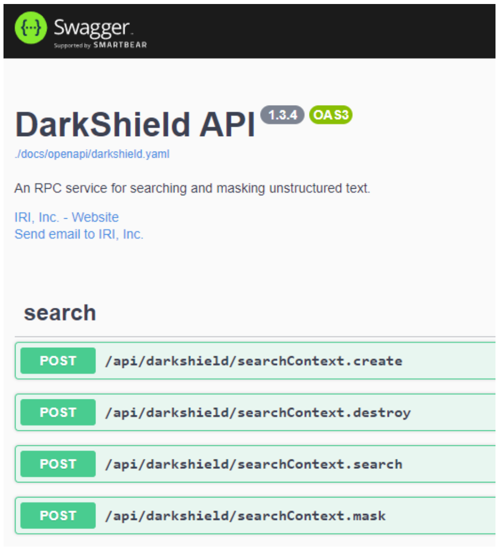


Command Prompt

```
(c) 2013 Microsoft Corporation. All rights reserved.

C:\Users\cosort>darkshield -h
Usage: darkshield [[Criteria Path [-r] [-t=Tessdata]] | [-d=Dark Data Path]]
                  [-hV] [-b=Binary Directory] [-e=Executable]
Command Line API for the DarkShield search and remediation engine.
  -b, --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.
  -V, --version             Print version information and exit.
Search Criteria Options      Criteria Path         Path to the Search Criteria c
onfiguration 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>
```

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. The API is the most flexible option, and the best way to run large jobs.



DarkShield API 1.3.4 OAS3

/docs/openapi/darkshield.yaml

An RPC service for searching and masking unstructured text.

IRI, Inc. - Website
Send email to IRI, Inc.

search

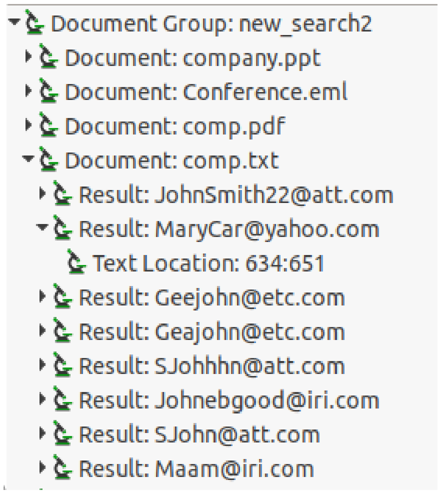
POST /api/darkshield/searchContext.create

POST /api/darkshield/searchContext.destroy

POST /api/darkshield/searchContext.search

POST /api/darkshield/searchContext.mask

Reporting and Using Results



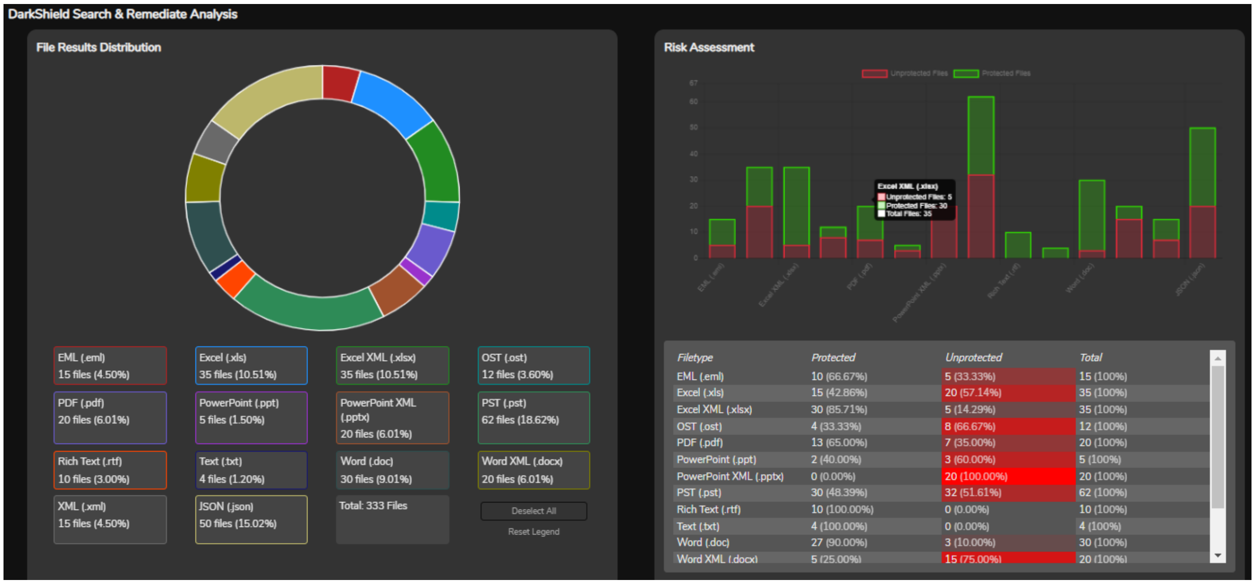
```
Document Group: new_search2
  Document: company.ppt
  Document: Conference.eml
  Document: comp.pdf
  Document: comp.txt
    Result: JohnSmith22@att.com
    Result: MaryCar@yahoo.com
      Text Location: 634:651
    Result: Geejohn@etc.com
    Result: Geajohn@etc.com
    Result: SJohhhn@att.com
    Result: Johnebgood@iri.com
    Result: SJohn@att.com
    Result: Maam@iri.com
```

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 Search & Remediate Analysis

File Results Distribution

EML (.eml) 15 files (4.50%)	Excel (.xls) 35 files (10.51%)	Excel XML (.xlsx) 35 files (10.51%)	OST (.ost) 12 files (3.60%)
PDF (.pdf) 20 files (6.01%)	PowerPoint (.ppt) 5 files (1.50%)	PowerPoint XML (.pptx) 20 files (6.01%)	PST (.pst) 62 files (18.62%)
Rich Text (.rtf) 10 files (3.00%)	Text (.txt) 4 files (1.20%)	Word (.doc) 30 files (9.01%)	Word XML (.docx) 20 files (6.01%)
XML (.xml) 15 files (4.50%)	JSON (.json) 50 files (15.02%)	Total: 333 Files	

Risk Assessment

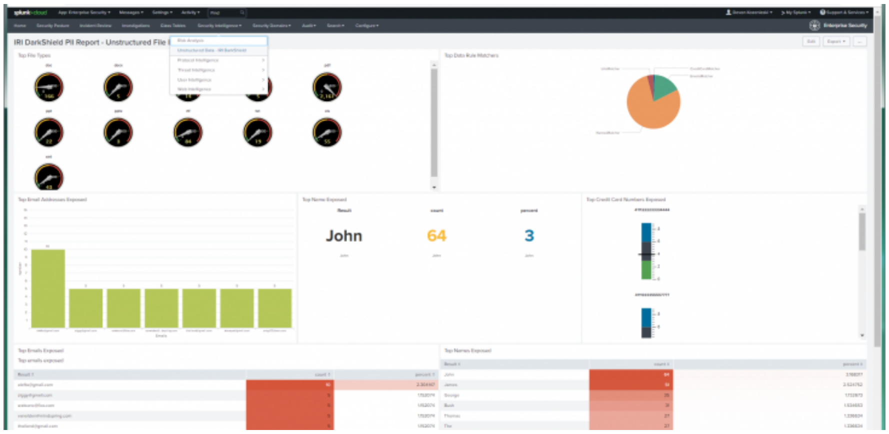
Filetype	Protected	Unprotected	Total
EML (.eml)	10 (66.67%)	5 (33.33%)	15 (100%)
Excel (.xls)	15 (42.86%)	20 (57.14%)	35 (100%)
Excel XML (.xlsx)	30 (85.71%)	5 (14.29%)	35 (100%)
OST (.ost)	4 (33.33%)	8 (66.67%)	12 (100%)
PDF (.pdf)	13 (65.00%)	7 (35.00%)	20 (100%)
PowerPoint (.ppt)	2 (40.00%)	3 (60.00%)	5 (100%)
PowerPoint XML (.pptx)	0 (0.00%)	20 (100.00%)	20 (100%)
PST (.pst)	30 (48.39%)	32 (51.61%)	62 (100%)
Rich Text (.rtf)	10 (100.00%)	0 (0.00%)	10 (100%)
Text (.txt)	4 (100.00%)	0 (0.00%)	4 (100%)
Word (.doc)	27 (90.00%)	3 (10.00%)	30 (100%)
Word XML (.docx)	5 (25.00%)	15 (75.00%)	20 (100%)

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

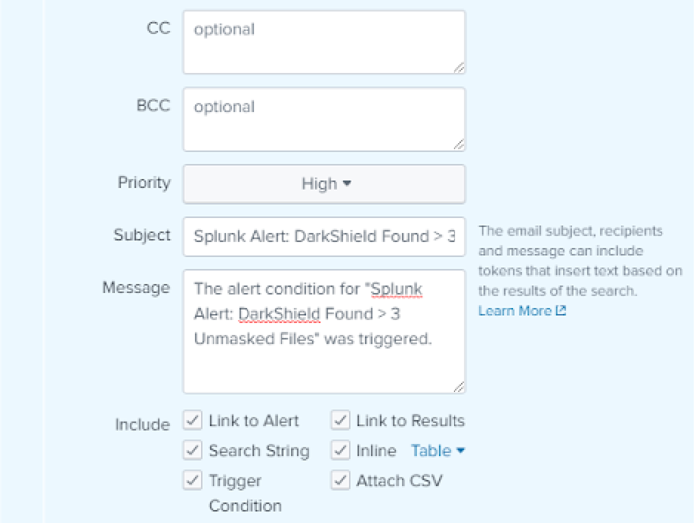
Security Information Event Management (SIEM) tools like Splunk Enterprise Security (ES) and modern analytic platforms like Datadog designed to create insights and enable actions from machine log data. As such, they can be used to categorize, graphically reveal, and report on security incidents from data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


IRI DarkShield PII Report - Unstructured File
Top File Types
Top Data Rule Matchers
Top Email Addresses Exposed
Top Name Exposed
Result
count
percent
John
64
3
Top Credit Card Numbers Exposed
Enterprise Security

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.



CC	optional
BCC	optional
Priority	High ▾
Subject	Splunk Alert: DarkShield Found > 3
Message	The alert condition for "Splunk Alert: DarkShield Found > 3 Unmasked Files" was triggered.
Include	☑ Link to Alert ☑ Link to Results ☑ Search String ☑ Inline Table ▾ ☑ Trigger Condition ☑ Attach CSV

The email subject, recipients and message can include tokens that insert text based on the results of the search. Learn More

As DarkShield searching and masking lobs 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 can find and mask PII in these file types:

Text	Documents	Images
.asc	.doc/x	.bmp
.html & .eml	.ppt/x	.gif
.fhir, .hl7 & .x12	.xls/x	.jpg/x/2
.json &.xml	.pdf	.png
.txt	.rtf *(search only)*	.tiff
.log.	Parquet	DICOM

and in these:

Data Silos & Databases

LAN, Related	Amazon	More Clouds/Apps	Additional Sources
Local & SMB	*CloudWatch*	*Box & SalesForce*	Couchbase, Redis, Solr
FTP/HTTP/MINA	DynamoDB	Elasticsearch	Cassandra, CosmosDB, MongoDB
Azure BLOB	RDS	*Facebook & LinkedIn*	Google BigTable & HBASE
GCP Storage	Redshift	*Google Apps*	JDBC (RDBs) *& JPA*
Sharepoint/OneDrive	S3 Buckets	*jclouds*	Kafka & *MQTT*

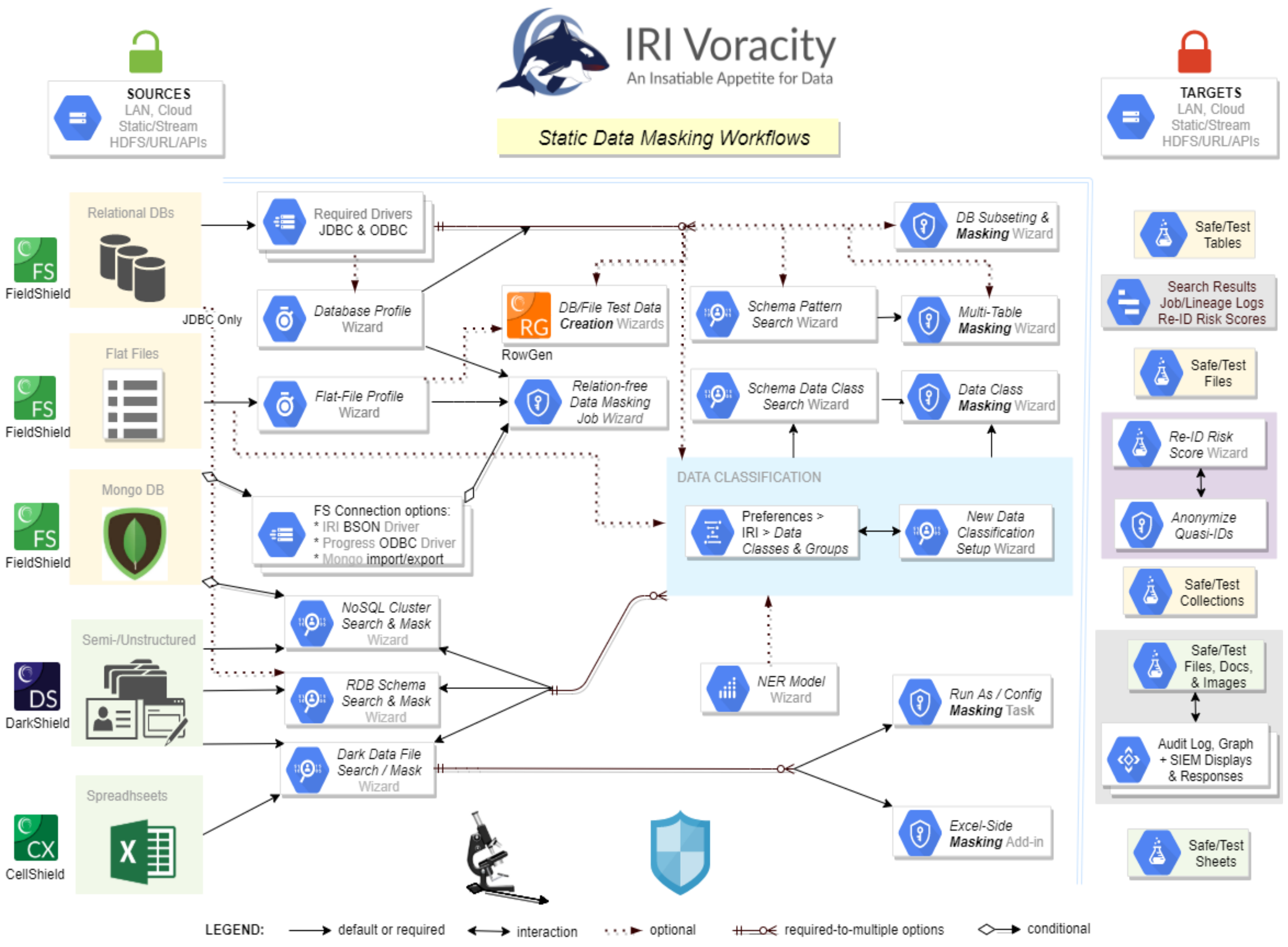
Sources in italics are on the support roadmap but are not yet enabled. If your file format or silo is not on the list above, please contact darkshield@iri.com to ask if it has been added since the publication of this booklet, or when it could be added.

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.

IRI Voracity
An Insatiable Appetite for Data



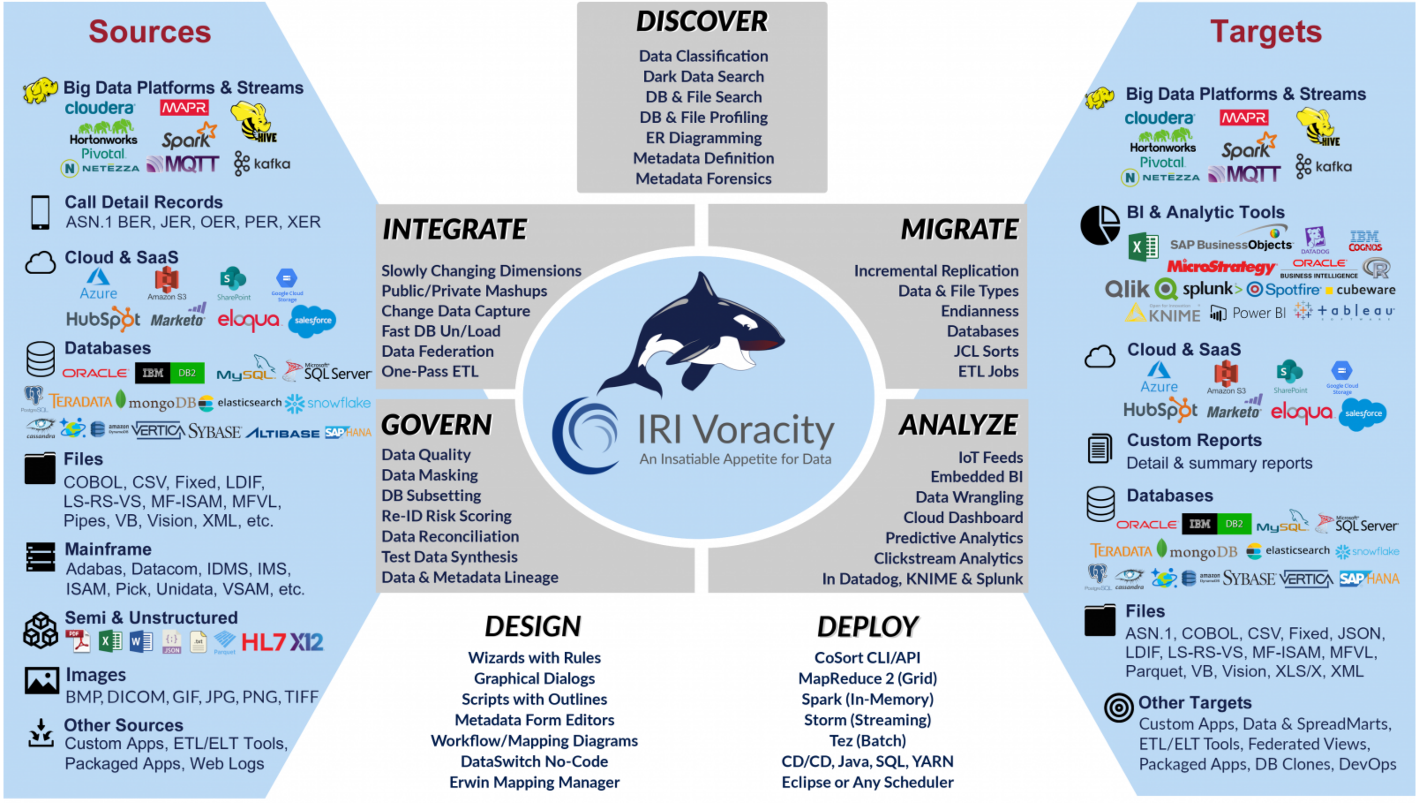
Static Data Masking Workflows

SOURCES
LAN, Cloud
Static/Stream
HDFS/URL/APIs

TARGETS
LAN, Cloud
Static/Stream
HDFS/URL/APIs

FieldShield
Relational DBs
JDBC Only
FieldShield
Flat Files
FieldShield
Mongo DB
DarkShield
Semi-/Unstructured
CellShield
Spreadhseets
Required Drivers JDBC & ODBC
Database Profile Wizard
Flat-File Profile Wizard
FS Connection options:
* IRI BSON Driver
* Progress ODBC Driver
* Mongo import/export
NoSQL Cluster Search & Mask Wizard
RDB Schema Search & Mask Wizard
Dark Data File Search / Mask Wizard
RG
DB/File Test Data Creation Wizards
RowGen
Relation-free Data Masking Job Wizard
DB Subseting & Masking Wizard
Schema Pattern Search Wizard
Multi-Table Masking Wizard
Schema Data Class Search Wizard
Data Class Masking Wizard
DATA CLASSIFICATION
Preferences > IRI > Data Classes & Groups
New Data Classification Setup Wizard
NER Model Wizard
Run As / Config Masking Task
Excel-Side Masking Add-in
Safe/Test Tables
Search Results Job/Lineage Logs Re-ID Risk Scores
Safe/Test Files
Re-ID Risk Score Wizard
Anonymize Quasi-IDs
Safe/Test Collections
Safe/Test Files, Docs, & Images
Audit Log, Graph + SIEM Displays & Responses
Safe/Test Sheets

LEGEND: ⟶ default or required ⟷ interaction ···▸ optional ‖—o< required-to-multiple options ◇⟶ conditional



Sources

Big Data Platforms & Streams

Call Detail Records
ASN.1 BER, JER, OER, PER, XER

Cloud & SaaS

Databases

Files
COBOL, CSV, Fixed, LDIF, LS-RS-VS, MF-ISAM, MFVL, Pipes, VB, Vision, XML, etc.

Mainframe
Adabas, Datacom, IDMS, IMS, ISAM, Pick, Unidata, VSAM, etc.

Semi & Unstructured

Images
BMP, DICOM, GIF, JPG, PNG, TIFF

Other Sources
Custom Apps, ETL/ELT Tools, Packaged Apps, Web Logs

DISCOVER

Data Classification
Dark Data Search
DB & File Search
DB & File Profiling
ER Diagramming
Metadata Definition
Metadata Forensics

INTEGRATE

Slowly Changing Dimensions
Public/Private Mashups
Change Data Capture
Fast DB Un/Load
Data Federation
One-Pass ETL

MIGRATE

Incremental Replication
Data & File Types
Endianness
Databases
JCL Sorts
ETL Jobs

IRI Voracity
An Insatiable Appetite for Data

GOVERN

Data Quality
Data Masking
DB Subsetting
Re-ID Risk Scoring
Data Reconciliation
Test Data Synthesis
Data & Metadata Lineage

ANALYZE

IoT Feeds
Embedded BI
Data Wrangling
Cloud Dashboard
Predictive Analytics
Clickstream Analytics
In Datadog, KNIME & Splunk

DESIGN

Wizards with Rules
Graphical Dialogs
Scripts with Outlines
Metadata Form Editors
Workflow/Mapping Diagrams
DataSwitch No-Code
Erwin Mapping Manager

DEPLOY

CoSort CLI/API
MapReduce 2 (Grid)
Spark (In-Memory)
Storm (Streaming)
Tez (Batch)
CD/CD, Java, SQL, YARN
Eclipse or Any Scheduler

Targets

Big Data Platforms & Streams

BI & Analytic Tools

Cloud & SaaS

Custom Reports
Detail & summary reports

Databases

Files
ASN.1, COBOL, CSV, Fixed, JSON, LDIF, LS-RS-VS, MF-ISAM, MFVL, Parquet, VB, Vision, XLS/X, XML

Other Targets
Custom Apps, Data & SpreadMarts, ETL/ELT Tools, Federated Views, Packaged Apps, DB Clones, DevOps

PartnerNet Silver Partner Novell
intel Software Partner
IBM Business Partner
Business Partner hp
DataSwitch Connecting Data
erwin
eclipse
msdn
MICRO FOCUS
melissa
redhat
ORACLE PARTNERNETWORK

IRI DarkShield
Unstructured Data Search & Security

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/darkshield

IRI
Total Data Management