Data X-Ray
Create order out of data chaos

Client Case Study: Automatic anonymising and desensitising of health and safety data

HSE
Health & Safety Executive
DISCOVERING SAFETY

HSE's anonymisation task was to take 12.5 person years in manually redacting 600,000 documents. The Data X-Ray reduced that to 1 machine day. The result? 4,500x less time at 49x less cost.

Challenge

The UK's Health and Safety Executive (HSE), the UK's Health and Safety regulator, was looking for innovative proposals for solutions that are able to accurately and expediently desensitise and anonymise health and safety information sources. This might include information held in both structured and unstructured data formats, contained in spreadsheets, databases and reports. The latter might include reports in Word documents, PDFs, or any other machine readable file formats, including both short (e.g. a few pages) as well as longer (>50 pages) volumes.

Solution

The solution called for a tool that is able to i) find where sensitive data is within large unstructured (and possibly structured) datasets and across various file formats, ii) flag where that particular data might be, and iii) automatically isolate sensitive data from the original dataset.

HSE selected the Data X-Ray, Ohalo's proprietary automated data mapping, search, and redaction tool, to carry out this work.

The Data X-Ray features connectors built for both unstructured and structured data sources in all principal file formats to i) extract data from its original format, ii) analyse that data to discover potentially sensitive tokens, iii) separate sensitive data from non-sensitive data, and iv) output that data to a usable format.

The solution involved a service installed within an HSE environment and provided to HSE and its partners to share and redact data with the goals of:

• Sharing data with research partners in a GDPR compliant way

Data X-Ray
Data-Redaction

Process at 100,000s of words per second
Different from a human, the Data XRay can read 100,000s of words per second, unlocking use cases not possible with manual processes

Achieve near-human anonymisation accuracy in unstructured data
Given enough training, the Data XRay can accurately redact up to 99% of records processed.

ohalo

https://www.ohalo.co | hello@ohalo.co

Data X-Ray
Data-Redaction

Having third party industry partners and other subject matter experts able to share data with HSE in a GDPR compliant way

Evaluating automatic basic document organization and categorization tasks with machine learning to prepare large data sets for more effective analysis in the future.

What was done?

Ohalo set up a Data X-Ray server in HSE's environment. The server ingested data from an HSE data source (HSE RIDDOR report data, which are an accident reporting format for HSE), analysed that data to identify sensitive personal data, and redacted it.

A team led by the HSE data science team worked with Ohalo to identify any sensitive information that had not been redacted and the ability to personally identify individuals or entities by 'joining up the dots', for instance by linking PII data to public data to infer identity.

Where false or true positives were discovered by the HSE team, the Data XRay enabled the team to update the models with example training information to facilitate improved redaction results. Multiple techniques were used to update the models, such as adding new ML classes, dictionaries, and regular expressions and data engineering techniques that involved ensuring unique word tokens are correctly combined to achieve better token concatenation.

Results

The effectiveness of Ohalo Data X-Ray anonymisation was evaluated against manual anonymisation. This evaluation was based on 1,998 RIDDOR reports used for the Construction Division RIDDOR dashboard (https://www.hse.gov.uk/construction-dashboard/) which was manually anonymised in 2017 and made public.

The standard of assessment used for significant breach is that of the UK's privacy regulator, the Information Commissioner's Office (ICO), and involved making a determination as to whether there will likely be a risk to people's rights and freedoms.

From the 1,998 RIDDOR reports analysed 743 contained sensitive information. Anonymisation using Ohalo Data X-Ray resulted in 94 retaining some form of PII of which 19 would be considered sufficient for a significant breach under the ICO guidance.

Of the 743 manually anonymised records containing personal data, 213 (29%) were identically auto-anonymised by Data XRay.

The Data X-Ray's automated anonymisation removed PII resulting in 99%+ of the records being anonymised.

The majority of the differences between manual and auto anonymisation were due to both under and over redaction. However, some minor differences are also due to manual alterations (e.g. spelling corrections) made to the original records during the anonymisation process and also some differences due to manual under redaction. Notably, 69 of the manually redacted records retained elements of 'sensitive' text e.g. tool brand names, first name, company acronyms, and motorway names. This serves to show that even manual redaction is not 100% consistent. In this case individual's assessments of sensitivity led to differing decisions with sensitive elements that were not PII.

The table below lists the categories and thenumber of records retaining sensitive textpost-anonymisation by class of data. (Notethat a single record may contain text frommultiple categories i.e. there is some overlap within the reported figures). Of the 1,998 total records, 94 records retained PII, and only 19 were considered to be a significant breach of GDPR, as defined by the method of assessment outlined above. This serves to show that the Data X-Ray's automated anonymization removed PII resulting in 99%+ of the records being anonymised.

Post-anonymisation results by retained data type (unredacted data) compared against 1,998 manual-redacted records

Category of under-redacted sensitive text	No. of records	Percentage remaining
PII (Significant Breach)	94 (19)	5% (1%)
Gender (Title)	48	2%
Company Name	42	2%
Location	83	4%
Date	34	2%
Reference numbers	31	2%

Going forward to drive better health and safety outcomes in the world

The engagement proved that the Data XRay successfully redacts personal and sensitive health and safety data from unstructured data sets to a very high degree of accuracy. Using the anonymised version of the data, HSE is able to share data with third party researchers for the furthering of HSE's mission to prevent death, injury and ill health to those at work and those affected by work activities. The end result is that HSE is able to unlock their very valuable data science personnel to drive value towards the analysis of the data rather than the engineering of the data.

Data X-Ray
Data-Redaction

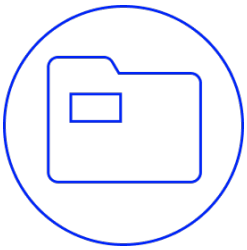
Data X-Ray for Anonymisation

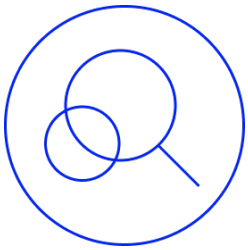
ENGINEERING YOUR PRIVATE DATA

Automated data ingestion and redaction at scale with the customisation required for your data

Integrate with native datasources or use Data XRays no-integration cloud options
Companies can operate literally thousands of databases. Data X-Ray integrates to most datasource types from SQL databases to Windows file servers to cloud storage in seconds. Data X-Ray also provides no-integration options using secure managed cloud storage.

Manage separate redaction projects with casefiles and custom classifiers
The Data X-Ray allows you to build out casefiles with subsets of data and overlay custom classifiers onto those subsets of data for ultimate accuracy. This is all available out-of-the-box on day one but can integrate to your existing data pipelines seamlessly through the REST API.

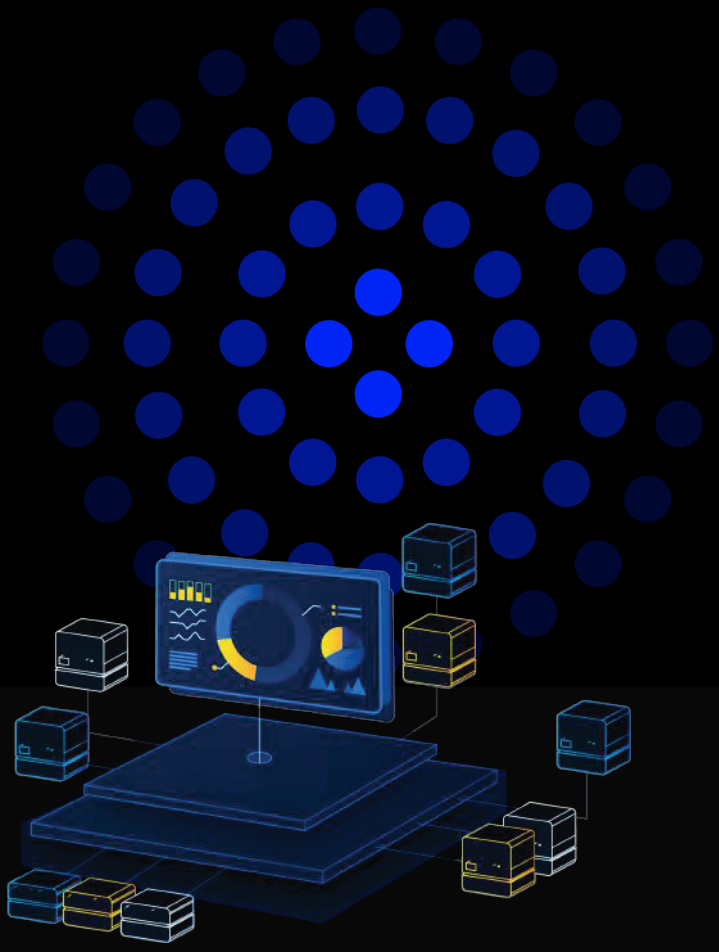




Build your own custom classes to find data that is sensitive in your context
The Data X-Ray rules customisation allows a user to define new AI classes, build regular expressions, and match keywords that are sensitive within a particular data pipeline context.

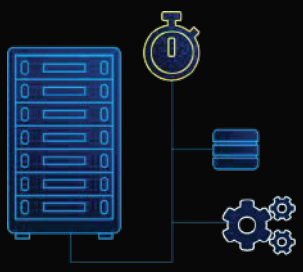
Data X-Ray
Create order out of data chaos

Security and Scalability

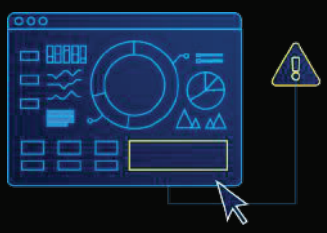


TOTAL SECURITY FOR YOUR DATA

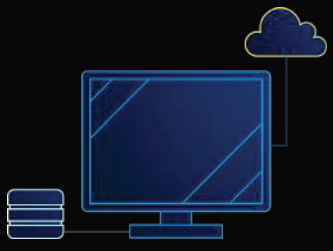
Customized networking and security, support for enterprise datasource types, and API access



On premise and private cloud support
A modern enterprise will have a diversity of datasources that power their bottom line. Ohalo's connectors support both cloud and on premise from a single solution so you can easily monitor your cloud object storage and your internal databases with a couple button clicks.



Support for enterprise grade datasources
Out of the box, the Data X-Ray supports the most common datasource types, with native support for SQL databases, One Drive, Windows file systems, and more. All deployed in your own environment so you can be sure your data is secured.



API support
No tool is an island. Comprehensive API support means that you can easily integrate the Data X-Ray results into existing data pipelines, data catalogs, document management systems, data visualization tools, and other internal and external systems that are key to your organization's operations.