Skip to main content

Navigation menu

Menu

Search GOV.UK

Home
Find out how algorithmic tools are used in public organisations

MoJ : Splink Master Record

Splink is an open-source tool for probabilistic data linkage tool that enables fast, accurate and scalable linkage and deduplication of data that lacks unique IDs

From:
Cabinet Office, Department for Science, Innovation and Technology and Government Digital Service
Published
6 October 2025

Organisation:
Ministry of Justice
Organisation type:
Ministerial department
Function:
Public order and safety
Capability:
Discovery
Task:
Clustering
Phase:
Production
Region:
Wales and England
Date published:
6 October 2025
ATRS version:
v4.0

Contents

1. Summary
1 - Name
2 - Description
3 - Website URL
4 - Contact email

2. Tier 2 - Owner and Responsibility
    1. 1.1 - Organisation or department
    2. 1.2 - Team
    3. 1.3 - Senior responsible owner
    4. 1.4 - Third party involvement
3. Tier 2 - Description and Rationale
    1. 2.1 - Detailed description
    2. 2.2 - Benefits
    3. 2.3 - Previous process
    4. 2.4 - Alternatives considered
4. Tier 2 - Deployment Context
    1. 3.1 - Integration into broader operational process
    2. 3.2 - Human review
    3. 3.3 - Frequency and scale of usage
    4. 3.4 - Required training
    5. 3.5 - Appeals and review
5. Tier 2 - Tool Specification
    1. 4.1.1 - System architecture
    2. 4.1.2 - System-level input
    3. 4.1.3 - System-level output
    4. 4.1.4 - Maintenance
    5. 4.1.5 - Models
6. Tier 2 - Model Specification
    1. 4.2.1. - Model name
    2. 4.2.2 - Model version
    3. 4.2.3 - Model task
    4. 4.2.4 - Model input
    5. 4.2.5 - Model output
    6. 4.2.6 - Model architecture
    7. 4.2.7 - Model performance
    8. 4.2.8 - Datasets and their purposes
7. 2.4.3. Development Data
    1. 4.3.1 - Development data description
    2. 4.3.2 - Data modality
    3. 4.3.3 - Data quantities
    4. 4.3.4 - Sensitive attributes
    5. 4.3.5 - Data completeness and representativeness
    6. 4.3.6 - Data cleaning
    7. 4.3.7 - Data collection
    8. 4.3.8 - Data access and storage
    9. 4.3.9 - Data sharing agreements
8. Tier 2 - Risks, Mitigations and Impact Assessments
    1. 5.1 - Impact assessments
    2. 5.2 - Risks and mitigations

1. Summary

1 - Name

Splink

2 - Description

Splink is an open source Python library for probabilistic record linkage. It is designed for linking and deduplicating datasets that lack a unique identifier

It is used to link persons between the Ministry of Justice's courts, prisons and probation datasets in both batch and real-time deployments. For example: - It is used weekly to refresh linked datasets for statistical analysis - It is used in courts to find probation records associated with individuals coming to court - It is being piloted as part of Core Person Record, a product that aims to create a unique identifier for persons across prisons, probation and the criminal courts - It is used to find Police National Computer (PNC) numbers associated with individuals, in order to request relevant arrest information from the police.

3 - Website URL

https://github.com/moj-analytical-services/splink

4 - Contact email

data_linking_team@justice.gov.uk

Tier 2 - Owner and Responsibility

1.1 - Organisation or department

Ministry of Justice

1.2 - Team

Data Linking team

1.3 - Senior responsible owner

Chief Data Scientist

1.4 - Third party involvement

No

Tier 2 - Description and Rationale

2.1 - Detailed description

Splink compares records of individuals within and across case management systems throughout the justice system (prison, probation and courts). When comparing the personal information (names, dates of birth, addresses etc.) of these individuals, the Splink model produces a probability score that the two records refer to the same person. This is known as probabilistic data linkage. Any record pairs with a match probability above a specified threshold are then considered as the same person, with this person being assigned a new linked identifier.

Details of the statistical model used can be found in the documentation website here: https://moj-analytical-services.github.io/splink/topic_guides/theory/fellegi_sunter.html

Splink is integrated into a number of systems. At the moment, it is used for:

- JustLink datasets: Each week, data is linked between prisons, probation and courts. The result is a fully anonymised lookup table that contains the links between records (but NOT the personal identifiers). The data is available to analysts to perform statistical analyses, such as understanding how long it takes cases to progress between courts.

- Probation in Court: To find probation records associated with an individual when they come to court, as part of the process of preparing a case for sentence.

- Core Person Record. This is a real-time linkage system that is currently being piloted that aims to create a unique identifier that links together each person across MoJ's criminal data systems (specifically courts, prisons and probation). In this system, as records are created and updated, Splink is used to predicts whether the record links to other existing records in the system.

- Police data sharing pilot. A regular data sharing system being piloted in Essex. Splink is used to identify Police National Computer (PNC) numbers associated with individuals supervised by North Essex Probation Delivery Unit. Those numbers are sent to the police each day to identify if any arrests have occurred.

Whilst the Splink algorithm uses personal identifiers to match records, it does not make personal identifiers more widely visible. Access to the PII itself remains limited to authorised staff under existing controls.

2.2 - Benefits

The introduction of Splink increases the speed and accuracy of linkage. The availability of linked data has a variety of benefits such as: - More reliable research and policy analysis – Analysts can now generate insights based on linked justice system data rather than fragmented datasets, supporting evidence-based decision-making. - Reduction in data duplication – By deduplicating records, Splink ensures that justice system statistics and analytics reflect unique individuals rather than multiple records of the same person. - Improved operational efficiency – The ability to track individuals across the justice system more accurately supports better resource allocation and case management. These improvements benefit the public by furthering the MoJ's priority outcomes—delivering swift access to justice, protecting the public, and reducing reoffending

2.3 - Previous process

Prior to the introduction of Splink, linking records across the justice system was done by different teams in different ways.

Linked data for analysts relied on a time consuming process of deterministic linkage. As a result, it was up to a year before linked data was available to analysts, slowing down the process of important work such as evaluating the effectiveness of interventions

For linking and deduplicating data between operational systems, a mixture of deterministic linkage and human intervention has been used.

2.4 - Alternatives considered

Before we built Splink, we tested several existing open-source packages (e.g. the R fastlink and the Python recordlinkage package) for suitability. None of the existing packages for probabilistic record linkage worked at the scale required for data linkage in government (10s-100s of millions of records). Therefore the Internal Data Linking team in the Ministry of Justice built Splink.

Deterministic (i.e. rules-based) linkage options were also considered. However, deterministic linkage was deemed unsuitable due to its: - Inability to capture nuance - Tendency for high false negative links - Difficulty in managing large numbers of rules for complex datasets.

Tier 2 - Deployment Context

3.1 - Integration into broader operational process

Splink is used to determine which records pertain to the same individual across a range of datasets.

The process of matching is usually automatic, but in some cases, Splink is used to show a list of possible matches to a human via a Graphical user interface (GUI).

The tool is capable of providing detailed information to the user on exactly why a particular prediction was made, see https://moj-analytical-services.github.io/splink/demos/tutorials/06_Visualising_predictions.html. This capability is predominantly used at the model design and training stage, as opposed to inference.

The group of records pertaining to an individual is called a ‘cluster’, and this is useful for both aggregate statistical use cases, such as identifying how long it takes for cases on average to progress through the court system, to operational use cases such as preparing a case that is being seen in the courts for sentence.

3.2 - Human review

Human review is typically used during the model training and quality assurance process to quantify accuracy. In use cases where Splink provides a list of possible matches to a human, the human then makes the final decision.

3.3 - Frequency and scale of usage

Splink is used continuously as new and updated data on persons in the justice system is received by the department.

3.4 - Required training

Not applicable - in most cases, the process of matching is automatic and happens in the background. Relevant training is in place for staff who perform manual merges and unmerges.

3.5 - Appeals and review

There are no complaint procedures specific to Splink itself since it does not directly make decisions about individuals. However, there is a formal complains procedure for the various parts of the justice system which could be used if something was perceived to have gone wrong. Where

linkage errors are identified, it is possible to record manual overrides in the system to prevent them re-occuring.

Tier 2 - Tool Specification

4.1.1 - System architecture

https://github.com/moj-analytical-services/splink

4.1.2 - System-level input

Tabular data containing personally identifiable information from the courts, prisons and probation system

4.1.3 - System-level output

Match scores for pairs of records. If the score exceeds a certain threshold, the model matches the records together. These predictions are then turned in to ‘clusters’ - groups of records from the various input data systems which pertain to the same individual. These clusters provide a succinct summary of the links found by the model.

4.1.4 - Maintenance

Continuous monitoring of matching results to identify errors. Re-training does not follow a regular schedule - instead it would be done as part of continuous improvement or in response to a rise in errors.

4.1.5 - Models

Splink uses a single model for linkage - the Fellegi-Sunter model. This model is based on Bayesian Statistics, is well-researched and understood as the industry standard for record linkage.

For more on how the Fellegi-Sunter algorithm works, see the Record Linkage Theory section of the Splink docs site (https://moj-analytical-services.github.io/splink/topic_guides/theory/fellegi_sunter.html) and the academic paper (https://imai.fas.harvard.edu/research/files/linkage.pdf) used as the basis for the implementation of the algorithm.

Tier 2 - Model Specification

4.2.1. - Model name

Splink

4.2.2 - Model version

Latest version as of June 2025 is Splink v4.0.8

4.2.3 - Model task

Data linking and deduplication, i.e. determining whether a pair of records pertains to the same individual or different individuals.

4.2.4 - Model input

Pairs of records containing personally identifiable information from the courts, prisons and probation system.

4.2.5 - Model output

Match scores for the pairs of records. If the score exceeds a certain threshold, the model matches the records together.

4.2.6 - Model architecture

Probabilistic Fellegi-Sunter record linkage model trained using unsupervised learning (the Expectation Maximisation (EM) algorithm). The EM approach is documented here: https://www.robinlinacre.com/em_intuition/ and a detailed tutorial of how model training works is provided here: https://moj-analytical-services.github.io/splink/demos/tutorials/00_Tutorial_Introduction.html

4.2.7 - Model performance

Data linkage is an unsupervised problem, so traditional machine learning accuracy metrics (e.g. precision, recall, F1 score) cannot be relied upon to reflect the true performance of a model.

Clerical labelling (i.e. manual labelling by a human) has been performed on a sample of record pairs to provide a reference point for results generated by the model. These labels cannot be considered as a ""ground truth"" (such as in a supervised problem), as a human cannot be sure if two records match or not. The results of clerical labelling vary depending on the person labelling the data. Instead, metrics derived from these labels provide a rough guide of whether the linkage matches what a person would expect.

Model performance is also assessed by spot-checking record pairs (https://moj-analytical-services.github.io/splink/topic_guides/evaluation/edge_overview.html#spot-checking-pairs-of-records), where the outcomes for different types of matches can be assessed against what a human would expect. This is generally targeted for records close to a linkage threshold (over which a link is deemed to be valid). Tools, such as the Comparison Viewer Dashboard (https://moj-analytical-services.github.io/splink/charts/comparison_viewer_dashboard.html), are provided within Splink to facilitate this exploration.

4.2.8 - Datasets and their purposes

DELIUS (probation) NOMIS (prison) Common Platform (criminal courts) LIBRA (magistrates' court)

2.4.3. Development Data

4.3.1 - Development data description

DELIUS (Probation) NOMIS (Prison) Common Platform (Criminal Courts) LIBRA (Magistrates' Court)

For JustLink only: FamilyMan (Family Courts) CaseMan (Civil Courts)

None of these dataset are open data

4.3.2 - Data modality

All datasets are tabular text data

4.3.3 - Data quantities

Size of data pre-deduplication is: LIBRA: 19.6m records Common Platform: 2.9m records NOMIS: 2.2m records DELIUS: 2.4m records Familyman: 20.9m records Caseman: 17.7m records

Data was not split for training

4.3.4 - Sensitive attributes

For most applications, core fields used are as follows: - Date of Birth - Name(s) - Current and past addresses - Sentence date(s)

In some applications such as deduplicating prison data, we also use additional fields: - Ethnicity - Birth place - Nationality - Gender - Person's height and weight

These attributes are to identify whether pairs of records pertain to the same person

4.3.5 - Data completeness and representativeness

The data is the full population of individuals in the justice system, not a sample.

Data suffers from a variety of data quality problems including missingness in, typos, and the use of aliases or incorrect data.

The purpose of Splink is to address these data quality problems.

4.3.6 - Data cleaning

Data cleaning and standardisation is performed to ensure records are comparable. The actions undertaken are for instance; upper casing, standardisation of punctuation, removal of invalid postcodes, standardising format of DoB to yyyy-mm-dd

4.3.7 - Data collection

Individuals' personally identifiable data is routinely collected in the department's administrative data systems for the purposes of managing court cases and offenders. A key reason the department and its agencies collects personal identifiers is to ensure we can be confident about the identity of individuals.

4.3.8 - Data access and storage

Development data is stored securely in the department's analytical platform https://user-guidance.analytical-platform.service.justice.gov.uk/ .

The Analytical Platform in hosted in a cloud-based ecosystem that is easy to access remotely from all MoJ IT systems. Designed for data at security classifications OFFICIAL and OFFICIAL-

SENSITIVE, we follow NCSC Cloud Security Principles, implementing features such as: - two-factor authentication - data encryption at rest and in transit - granular access control - extensive tracking of user behaviour, user privilege requests/changes and data flows - multiple isolation levels between users and system components

The data is accessible only to staff on the data linking team who need access for model development and quality assurance. There is no de-identification because the personal identifiers are essential to the task of predicting whether two records pertain to the same person.

4.3.9 - Data sharing agreements

N/A - all data is internal to MoJ

Tier 2 - Risks, Mitigations and Impact Assessments

5.1 - Impact assessments

- DPIAs/DPIA screenings conducted for each Splink use case
- Significant work into bias in data linkage: https://moj-analytical-services.github.io/splink/blog/2024/08/19/bias-in-data-linking.html https://moj-analytical-services.github.io/splink/blog/2024/12/02/bias-in-data-linking-continued.html
- For Core Person Record, the newest application of Splink, MoJ's new ethics framework is being followed (https://www.gov.uk/government/publications/ministry-of-justice-ai-and-data-science-ethics-framework)

5.2 - Risks and mitigations

Key risk 1: Errors can be made in data linkage.

Detailed clerical review has been used to mitigate this risk and attempt to quantify error rates. Given existing approaches to the problem of data linkage, we believe the use of Splink reduces (but does not totally eliminate) errors and so reduces the risks associated with linkage errors.

We have automated processes in place to monitor linkage quality and detect anomalies, and are working to improve these processes to more quickly identify and rectify errors.

Key risk 2: Linkage may be more or less accurate for different groups. This could occur, for instance, if more typos are made for records for people of a certain demographic group, or there is greater missingness in the data for certain groups. Work has been undertaken to understand bias in data linkage. Again, given existing approaches to the problem of data linkage, we do not believe this problem is made worse by the use of Splink

Mitigation Differential linkage rates occur when we have poor data quality on peoples' identity and the primary purpose of the Core Person Record project is to improve data quality across all records, particularly those that are currently difficult to link.

Updates to this page

Published 6 October 2025

↑ Contents

Help us improve GOV.UK

To help us improve GOV.UK, we'd like to know more about your visit today. Please fill in this survey (opens in a new tab and requires JavaScript).

Cancel

Services and information

Benefits
Births, death, marriages and care
Business and self-employed
Childcare and parenting
Citizenship and living in the UK
Crime, justice and the law
Disabled people
Driving and transport
Education and learning
Employing people
Environment and countryside
Housing and local services
Money and tax
Passports, travel and living abroad
Visas and immigration
Working, jobs and pensions

Government activity

Departments
News
Guidance and regulation
Research and statistics
Policy papers and consultations
Transparency
How government works
Get involved

Support links

Help
Privacy
Cookies
Accessibility statement
Contact
Terms and conditions
Rhestr o Wasanaethau Cymraeg
Government Digital Service