Skip to main content

Navigation menu

Menu

Search GOV.UK

Home
Find out how algorithmic tools are used in public organisations

MoJ: Data First (Splink)

Splink is an open-source tool for probabilistic data linkage developed for the Data First program. Data First provides academic researchers with datasets across the Justice system where individuals can be identified with a single, universal ID across domains.

From:
Cabinet Office, Department for Science, Innovation and Technology and Government Digital Service
Published
17 December 2024

Organisation:
Ministry of Justice
Organisation type:
Ministerial department
Function:
Defence
Capability:
Discovery and Management
Phase:
Production
Region:
Wales and England
Date published:
17 December 2024
ATRS version:
v3.0

Contents

Tier 1 Information

Name

Data First (Splink)

Description

Data First provides academic researchers with datasets across the Justice system (courts, prisons and probation) where individuals can be identified with a single, universal ID. Each dataset is deduplicated, giving a new ID for individuals within each domain, and then linked across domains to provide a cross-justice system linked ID. These IDs have been generated algorithmically using Splink, an open-source tool for probabilistic data linkage.

By creating a common ID across domains, researchers can perform unprecedented analysis of repeat users of the same services and journeys/cross-cutting analysis across multiple services.

Website URL

https://www.adruk.org/our-work/browse-all-projects/data-first-harnessing-the-potential-of-linked-administrative-data-for-the-justice-system-169/

https://www.gov.uk/guidance/ministry-of-justice-data-first

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

Contact email

datafirst@justice.gov.uk

Tier 2 - Owner and Responsibility

1.1 - Organisation or department

Ministry of Justice

1.2 - Team

Data First

1.3 - Senior responsible owner

Head of Data First

1.4 - External supplier 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.

For research in a single domain, this process is performed within each dataset to deduplicate individuals. For cross-domain analysis, a further linkage process creates a whole justice system linked identifier.

The deduplicated identifiers have been added to each of the (domain-specific) Data First datasets and these are used in combination with the XJS (Cross Justice System) table which acts as a lookup table between the individual system IDs, deduplicated ID and cross-domain linked IDs.

2.2 - Scope

The Data First datasets have been created solely for use by accredited researchers to perform cross-cutting analysis in the justice space.

The datasets provided are at an individual record (i.e. person) level and is anonymised for data protection purposes. This high level of granularity allows for more robust and flexible analysis.

The datasets are intended for identifying trends across the justice system to inform government policy, not for operational decision-making. A summary of the latest research can be found in the latest Data First Research Bulletin.

2.3 - Benefit

No cross-justice system linked datasets have been available to researchers previously. Some datasets have overlapping identifiers, but these are not populated consistently nor considered to be reliable enough for linkage at scale. As such, the benefit of these datasets comes from the new research opportunities that are now possible.

Examples of such research can be found on the Data First page on gov.uk.

2.4 - Previous process

N/A

2.5 - Alternatives considered

When scoping the Data First scheme, several existing open-source packages (e.g. fastlink) were tested for suitability. None of the existing packages for probabilistic record linkage worked at the scale required for data linkage in government (10s-100s 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 - Decision making Process

3.1 - Process integration

The Data First datasets are not integrated into a decision-making process.

They are datasets provided to academic researchers to study and identify trends within the justice system. Any resulting research may be used to build an evidence base to guide future policy decisions. In this instance, we strongly recommend that any analytical caveats from research include those provided by Data First for the linkage process.

3.2 - Provided information

The tool (i.e. new linked datasets) combines existing justice datasets with an additional identifier to link individuals across the justice system. This allows researchers to perform analysis across multiple domains.

3.3 - Frequency and scale of usage

At the time of publishing, Data First has facilitated around 40 projects that align with the department's strategic evidence priorities, as set out in the MoJ Areas of Research Interest.

Data linkage pipelines are run on weekly to provide linked data for internal analysis within the Ministry of Justice. The Data First datasets are republished periodically to provide the most recent data to researchers using the latest version of the linked data. This provides regular opportunities to improve and assess the linkage models between releases for Data First.

3.4 - Human decisions and review

The linkage feeding into the Data First datasets goes through several review process before being released to researchers.

For each release of data, a sample of records and person clusters are manually reviewed to ensure the models have performed as expected.

An overview of the high-level strategy we recommend for evaluating data linkage can be found in the Splink documentation and more detail on model performance for Data First specifically can be found in section 4.2.7 of this report.

3.5 - Required training

To access the datasets researchers must apply to become an accredited researcher through the ONS accredited researcher scheme. This includes training from the ONS' Research Services and Data Access team.

For more detailed information, please refer to the Data First Introductory User Guide. This document provides an overview of the potential of the linked data (Section 1), the limitations of data linkage (Section 2), and the requirements for getting access approved for a given researcher and or research project (Section 3).

3.6 - Appeals and review

Given researchers are not provided with personally identifying information, it is not expected that they will be able to assess whether a link between individuals is correct or not. However, as the models are regularly being reviewed within the MoJ, researchers can provide feedback if they believe they have found an error or anomaly which will be addressed in the next model iteration cycle.

Tier 2 - Tool Specification

4.1.1 - System architecture

The Data First datasets are made up of 43 different tables. These tables are structured across the 7 domains:

- Prisons (3 tables)
- Probation (5 tables)
- Criminal Courts (12 tables)
- Family Courts (3 tables)
- Civil Courts (12 tables)
- Offender Assessments (14 tables)
- Cross Justice Linkage (1 table)

The individual datasets are described in detail in the associated data catalogues.

The domain-specific, deduplicated identifiers within datasets are created by feeding each source dataset into Splink. Splink identifies links within the dataset (i.e. pairs of records with a match probability above a specified threshold) and the linked records are grouped and given a deduplicated identifier. These identifiers are added to the domain-specific datasets. For cross-domain linkage, all datasets are fed into Splink together to identify links. Once these links have been created, all records with a sufficiently high match probability are grouped and given a new linked identifier. These identifiers are added to the Cross Justice Linkage table.

The resulting Data First datasets, including deduplicated and linked identifiers, are transferred from the Ministry of Justice to the Office for National Statistics Secure Research Service and the SAIL Databank where researchers can apply to access them.

4.1.2 - Phase

Production

4.1.3 - Maintenance

The Data First Datasets are refreshed with new data and shared periodically. The datasets are reviewed before publication.

The linkage models used to produce the Data First outputs are used in the Ministry of Justice for internal analysis and are therefore being actively maintained and developed by the Internal Data Linking team.

4.1.4 - Models

The linkage underpinning the Data First datasets is performed by the Fellegi-Sunter model. The Fellegi-Sunter 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 and the academic paper used as the basis for the implementation of the algorithm.

Tier 2 - Model Specification

4.2.1 - Model name

Cross Justice System (XJS) Data First Splink Model

4.2.2 - Model version

2023-03-01 00:00:00

4.2.3 - Model task

Provide a similarity score between records of individuals across each of the justice datasets specified below (4.2.8). Records considered sufficiently similar (i.e. pass a given threshold) are clustered together.

4.2.4 - Model input

Person level data with personal identifiers (e.g. names, date of birth, address) for linkage purposes.

4.2.5 - Model output

Person level data with a new linked identifier column, where a records referring to the same individual have the same linked identifier.

4.2.6 - Model architecture

There are two layers of models in the XJS Data First Splink Model. For each dataset, a deduplication model is trained and links are generated. This identifies duplicate records within the same system (e.g. prisons). The second layer trains a model across all datasets to generate links between datasets (e.g. between prisons and probation). Each of these models does the same thing (i.e. generates links between pairs of records) but they are trained on a specific dataset or combination of datasets to produce optimal results. In total, there are 8 individual linkage models underlying the final linked data.

Each model is based on the relevant personal information available within each dataset (name, date of birth, internal ID numbers etc.).

Once the links have been generated from the deduplication and cross-dataset linkage models, a connected components algorithm clusters records together. Each cluster is then considered to be a single individual and given a new linked identifier.

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, 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, are provided within Splink to facilitate this exploration.

Work is ongoing to identify potential sources of bias in the underlying data linkage models (e.g. ethnicity, gender, non-anglicised names), including a dedicated internship programme in conjunction with the Alan Turing Institute.

4.2.8 - Datasets

Datasets included in the Data First products are:

- Magistrates' Court
- Crown Court
- Family Court
- Civil Court
- Prisons
- Probation
- Offender Assessment

All datasets that are linked as a part of Data First are included in the Data First Data Catalogues.

4.2.9 - Dataset purposes

The personally identifying information (PII) from each dataset is used for all parts of the linkage process (model training, validation, prediction).

All of this personal information is removed for the final output datasets.

Tier 2 - Risks, Mitigations and Impact Assessments

5.1 - Impact assessment

A Data Protection Impact Assessment (DPIA) has been completed for the Data First Project, including the creation and sharing of the Data First datasets. This Impact Assessment is developed and agreed upon with data owners before sharing the datasets with the trusted research environments.

Each time a researcher wants to use these datasets they are required to complete a research proposal which is scrutinised by various panels, including Data First staff and representatives of the data owner. These reviews ensure the proposal is feasible, addresses evidence gaps and identifies potential ethical issues. A further review is conducted of the outputs produced to ensure they are in line with the originally agreed scope and that the data has not been misused in any way.

Additional information can be found in the Data First Privacy and Data Protection guidance.

5.2 - Risks and mitigations

Data First uses the ONS's "Five Safes" framework to ensure information is kept safe and secure: - Safe People - Trained and accredited researchers are trusted to use data appropriately. - Safe Projects - Data are only used for valuable, ethical research that delivers clear public benefits. - Safe Settings - Access to data is only possible using secure technology systems. - Safe Outputs - All research outputs are checked to ensure they cannot identify data subjects. - Safe Data - Researchers can only use data that have been de-identified

More detail on this approach can be found in the Data First Privacy and Data Protection guidance.

Risks specific to Data First which have been mitigated include: Errors in the linkage - False Positive links (i.e. linking records referring to two different individuals) and False Negative links (i.e. not linking records referring to the same person). Depending on the researcher's question, the risk associated with one of these error types can be more important than the other. To mitigate these potential risks, the Data First datasets include linkages at three different levels of confidence thresholds (high, medium and low). The high threshold allows for fewer links (minimising False Positives) the low threshold allows for less confident links (minimising False Negatives), and the medium threshold provides a middle ground between the two. Providing multiple levels allows researchers to decide how they mitigate risks in the linkage based on the analysis being performed.

Misuse of Data - Data and any subsequent analysis is performed in a secure setting without internet access. This minimises the risk of identification as researchers are unable to combine data with external information to infer information about individuals. - Activity of researchers on the ONS Secure Research Service is closely monitored and access will be revoked if they go beyond the permitted uses of the data. - All outputs are checked to ensure they remain within the scope of the approved project.

Anonymisation - Data is only shared once all identifiers have been removed, and any additional personal descriptors (e.g. sex, gender, age) are not shared by default to minimise the risk of identifying individuals within the datasets. Some personal descriptors may be shared with researchers, if there is sufficient justification for the specific research question and the combination of descriptors will not risk identifying any individuals. Such requests are dealt with on a case-by-case basis. - Attempts to identify individuals are strictly against the terms and conditions to which researchers must agree.

Updates to this page

Published 17 December 2024

↑ Contents

Help us improve GOV.UK

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