Data First:

An Introductory User Guide

Harnessing the potential of linked administrative data for the justice system.

Version 8.1, September 2024

Contents

What is this guide for?

This guide has been developed to provide users and potential users with background information on the Data First programme, and to aid interested parties with initial enquiries through to submitting applications and accessing shared data. Through this guide, a reader should be able to find information on the overall aims of the programme, understand the key information contained within each shared dataset and access the relevant metadata. The guide should also provide an understanding of the different application pathways for accessing Data First datasets.

Section 1 provides an overview of the Data First programme, its aims, and the research potential of linked administrative data from the justice system. This also introduces the six main areas of work that make up the Data First programme, and the benefits of working with Data First for those both inside and outside of government. The information provided here will be useful for those who are new to the programme, or in the early stages of developing research projects that may benefit from working with Data First.

Section 2 covers the content of the available Data First datasets in more detail, including key variables, potential research questions and what approval(s) are required in order to access the data. This section can also signpost potential users to metadata on each of the datasets, which can help with shaping research questions and applications, and includes information on how data privacy, security and de-identification of the data are ensured. Information in this section can help with developing research questions and determining whether a research project would benefit from access to Data First datasets.

Section 3 covers the practicalities of gaining access to Data First datasets. This section will aid potential applicants in selecting where they wish to apply, how to complete the Office for National Statistics (ONS) accredited researcher training, and where users will be able to access their data and the resources available to them for analysis. This section is designed to help researchers who are familiar with the Data First programme, and know what information they would like to access, but are unsure which of the pathways to securing access to the data would be most beneficial.

Section 1: What is the Data First programme?

Programme overview

Data First is an ambitious data-linking, academic engagement and research programme led by the Ministry of Justice (MoJ) and funded by ADR UK (Administrative Data Research UK), who in turn are funded by the Economic Social Research Council (ESRC). The programme began in 2019 and has received a three-year extension to 2025.

Figure 1 - The link between different organisations in the setup of Data First

Data First unlocks the potential of the wealth of data already collected by MoJ by linking administrative datasets from across the justice system and beyond. It enables accredited researchers across government and academia to access anonymised, research-ready datasets ethically and responsibly. The programme also enhances the linking of justice data with other government departments (OGDs).

Data First datasets are deidentified, deduplicated and then shared, in most cases, with both the ONS Secure Research Service (SRS) and SAIL (Secure Anonymised Information Linkage) Databank (there are some exceptions to this, which can be found in Section 2, Table 1, where a full list of available datasets can also be found). Approved researchers can then access the relevant data through a number of methods, including ONS safe rooms, approved remote desktop connections and the Safe Pod Network. Data cannot be removed to be stored, for example, on a researcher's own PC or university servers.

By working in partnership with academics, OGDs, and third sector organisations to facilitate research in the justice space, the Data First programme is enabling analysis of user journeys, interactions, and outcomes across the justice system and with a range of other public services. This provides evidence to underpin the development of government policy and address social and justice issues.

Data First forms an integral part of MoJ's wider ambitions to enhance the way data and evidence is used to shape decision-making and drive improvements to justice outcomes. A more comprehensive, dedicated and coordinated approach to engagement with external partners, underpinned by the department's Areas of Research Interest 2020 (ARI), is key to achieving this. The linked administrative data made accessible via Data First enables some of the critical evidence gaps outlined in the ARI to be explored in collaboration with our academic partners. In doing so, the strategic research capabilities are strengthened

across government and academia to reinforce the impact of evidence at all stages of policy development and evaluation.

What is the potential of this newly linked data?

By linking administrative datasets across the whole justice system, the Data First programme is helping to build a picture of justice system users and interactions over time across the courts, prison and probation services. Understanding these characteristics, patterns of frequent use, and common transitions between different services can help develop our understanding of what works, and where improvements may be needed to inform government policies and services.

There is significant potential in these newly linked datasets. Data First is designed to facilitate links with our research and academic partners; by working in collaboration, priority areas for analysis can be identified to make best use of the data. The Approved external data requests log provides information on projects that have been granted permission to access Data First datasets, while internal research outputs that have been produced as a result of Data First can be found here.

Further linking of MoJ data with that of OGDs, such as work already done with the Department for Education (DfE), will enhance our understanding of how justice system users interact with other public services, and their needs, pathways and outcomes across a range of important measures.

What does Data First involve?

Data First comprises a number of elements, as outlined in Figure 2.

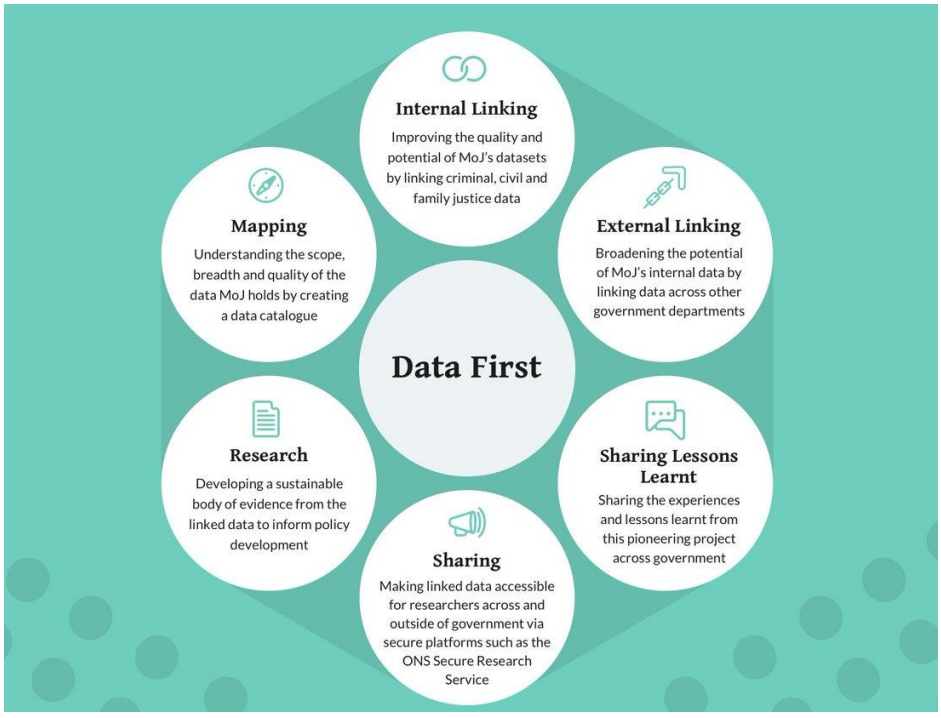

Figure 2 - Representation of the different workstreams within Data First

Interdisciplinary teams of data scientists, data engineers, statisticians and social researchers are leading the different workstreams of the Data First programme within MoJ:

- **Internal data linking** – development of a robust, automated linking pipeline between MoJ cross-justice system datasets using the Splink package.
- **External data linking** – establishing data shares with external partners, linking justice data with OGDs and managing research and analysis using the MoJ-DfE data share.
- **Data engineering and data mapping** – creating pipelines to bring new data sources into scope, agreeing what can be shared, designing and developing research-ready datasets with documentation.
- **Research, academic engagement and communications** – facilitating the link between Data First and the academic research community, supporting researchers to access and understand the data, and working in partnership to identify priority research questions to make best use of the linked datasets.

What are the benefits of Data First?

The work delivered across Data First offers a wide range of benefits, as summarised in Figure 3.

Figure 3 - Benefits of Data First

- **Research** – the programme enhances the strategic research capabilities of researchers across both government and academia.
- **Better understanding of justice system users** – the linkage of data provides a better understanding of justice system users, their characteristics and the journeys they take through the system.

- **Improving the evidence base** – linking justice-related datasets provides researchers with the capability to address research questions and develop a stronger evidence base to inform policy development and the effects of policy interventions.
- **Relationship with academia** – Data First brings together academia and government, building a partnership to utilise knowledge and expertise, and maximise the impact of the research.
- **Lessons learnt** – the programme allows MoJ to share the lessons they learn across government, academia and other stakeholders.
- **Improving the transparency of policy-making** – research published under the Data First programme further enhances the openness of the use of evidence in policymaking.

Section 2: Data shared through Data First

Where does the data come from?

The Data First programme links and shares extracts of data from a range of administrative databases owned by the Ministry of Justice and its agencies, His Majesty's Courts and Tribunals Service (HMCTS) and His Majesty's Prison and Probation Service (HMPPS), as well as works to put agreements in place with other government departments to share data from other sources. Currently, the datasets linked and shared by Data First include information on:

- Criminal court cases and defendants in magistrates' courts and the Crown Court in England and Wales
- Offenders in prison custody or under supervision of probation services in England and Wales
- Civil and family court jurisdiction cases and parties (the people and organisations involved) in England and Wales
- Department for Education data on young people in England's education data from the National Pupil Database (NPD) linked to Police National Computer (PNC) data on criminal histories.

In future, data will be made available on offender assessments and from other government departments subject to agreements. The justice sector has a complex landscape of management information systems. These have been developed at different points in time to meet the challenges of diverse tasks such as handling a child supervision order; processing a civil monetary claim; managing a criminal trial; assessing offender needs; or managing an offender's time in custody. Only data that are already being collected and stored on certain digital systems are in scope for inclusion in the programme. Many court processes are handled locally by individual courts, they frequently involve paper forms, documents, evidence and judgments that are not recorded centrally. The criminal justice system also involves data collection outside of the MoJ, including by police forces, the Home Office and the Crown Prosecution Service (CPS), which means that MoJ holds little information about alleged offences that do not reach the courts.

A number of changes to MoJ management information systems which affect how data is collected (for example the Court Reform Programme) are either planned or already underway. Data First is focusing on 'legacy' data systems which provide a number of years of useable data for research. Sometimes, this means that the most recent data ceases to be complete in coverage. Documentation will communicate this wherever possible.

How do Data First data compare to that reported elsewhere?

The Data First process includes work on structuring datasets for sharing and linking. While the same underlying administrative data sources may be used elsewhere (for example, in Official Statistics or previous research reports) differences are expected due to separate processing, especially where matching between multiple datasets has taken place. While material differences in trends or conclusions are not anticipated, researchers should be aware that analysis carried out using Data First datasets may not be exactly comparable to other published statistics or research.

Understanding the data available

- See summary of datasets available
- Full metadata is available from our data catalogues on the GOV.UK website.

The data catalogues for Data First include information about the coverage and source of the dataset as a whole; the structure of the dataset; and provide metadata for each variable shared, including their quality, format, type and description. They also include lists of values and lookups for categorical variables.

Researchers can use a data catalogue to understand what potential research questions a dataset can answer, and identify the variable names required.

You can also find searchable catalogues including Data First datasets from:

- ADR UK Data Catalogue
- ONS Secure Research Service metadata catalogue
- SAIL Databank Health Data Research Innovation Gateway

Using administrative data for research

What is administrative data?

Administrative data refers to information which was originally collected for operational purposes, rather than statistics or research, such as to enable the delivery of a public programme or service or to maintain records. The administrative data that is being used by Data First was originally collected to administer the justice system (or other government services), such as to process court cases or run prison and probation services day-to-day.

Despite research needs not generally being part of the collection design, administrative data sources can be a rich source of information for quantitative analysis and evaluation without imposing an additional burden on data subjects or costs to data controllers. By re-using administrative data and making it securely and ethically available to those who can make best use of it, its value for research in the public good can be maximised.

What can the data be used for?

Administrative data can help to answer certain types of questions. Broadly speaking, Data First can address questions about who is interacting with the justice system in what ways, the processes and timings involved, and 'hard' outcomes such as sentencing or frequency of repeat appearances. The kind of data shared through Data First cannot provide insight into unmet needs, circumstances and experiences of people engaging with justice services.

Data First focuses on justice system users (for example defendants, prisoners, parties to family court and civil court cases). There is little information about the judiciary, court and prison staff, or legal professionals. Additionally, limited administrative data is held by MoJ on some public users such as victims and witnesses.

Administrative data provides a good window into the big picture and small subsets. The datasets contain information on all relevant cases in England & Wales over the time periods covered. This means it is often possible to look at niche groups such as outcomes for different ethnicities, or at specific offence types.

Data quality issues

Data First identify appropriate extracts of data to make available for research purposes from sometimes complex existing pipelines or from new copies of live databases. While certain fields are used by Ministry of Justice analysts, for example for release in Official Statistics publications, the programme includes data for which less is known about quality and limitations. Documentation may also be incomplete. In releasing this resource to the wider research community, our understanding of the source data will be increased. By collaborating, sharing experiences and expertise our assessment of its strengths and weaknesses in addressing research questions can be improved.

Data collection is carried out largely by frontline staff and users of services. Gaps in coverage are to be expected where items have not been considered essential to the original operational need, or even where entire populations of interest or categories of experience are absent from administrative sources. Data may also be recorded inconsistently over time, across the country, or depending on entry method, meaning that data may not be of the desired quality or comprehensiveness to address important research questions. There is little harmonisation of the fields collected on different systems, creating inconsistency in data definitions, data formats and values. Researchers will therefore need to carry out their own cleaning and preparation ahead of analysis.

Over time, changes to management information systems, processes and policies may have introduced breaks in time series that could affect analysis and interpretation. While data linkage creates great potential for in-depth longitudinal questions to be considered, researchers should remain aware of the scope and origin of the datasets being made available through Data First.

Data linkage

What data is being linked?

The internal administrative datasets that MoJ are bringing together as part of Data First each represent an interaction of a 'user' of the justice system (for example, a defendant, offender, or a user of the civil or family courts) with justice processes or services.

Most datasets contain duplicates of individuals (i.e., many records pertaining to a single person), and the same individual may appear in different datasets (for example, both as a defendant in a criminal trial, and as a respondent in a family law case). The challenge is that generally no reliable unique ID exists, either within or between datasets, to link information about a person back to previous 'journeys' through the justice system.

Data First aims to provide a unique ID for researchers working with one dataset, and also to identify through linked datasets when the same individual is thought to appear across multiple datasets (for instance, in both a court and a prisons dataset).

Data First are linking records only at the level of an individual, to allow analysis of a person's journey through the justice system. Data First are not identifying networks of individuals linked by personal information such as shared addresses over time (although some relationships such as between co-defendants in a single criminal case are linked in the source data). Combined magistrates' courts and Crown Court datasets can also be linked at a case level, to follow the progression of a case through the court system.

Data linking process

Without a unique personal identifier, other identifying information is compared, such as names, date of birth and addresses, that are held in the source data to inform these decisions. This personal information will not be shared by MoJ and will be replaced in the data linking process by a meaningless identifier (one that has been generated for these datasets and is not used in any existing operational systems).

In some cases, two records will contain the same values in each of these fields, making it clear that they refer to the same defendant. However, duplicate records may not match for a variety of reasons, including:

- Typographical/phonetic errors
- Change of name/address
- Aliases/nicknames/diminutives
- Missing data

Internal data linking for Data First is done by adopting a probabilistic approach (using the canonical model of Fellegi and Sunter, 1969[1]) whereby each pair of records is assigned a match score based on the level of agreement in each attribute used for linking. Each attribute is assigned a weight that contributes to this match score, so a match on date of

[1] Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64(328), pp.1183–1210.

birth, for example, would influence our decision more than a match on gender, which adds less information.

Splink

To link the datasets required for Data First, a solution that can perform the necessary probabilistic data linkage at large scales has been developed by data scientists at MoJ.

The open-source Splink package uses Python and Apache Spark to link and deduplicate data flexibly, transparently and efficiently. The package implements the Fellegi-Sunter linkage model, estimating the parameters using the Expectation-Maximisation algorithm described by the authors of the fastLink R package in their paper.

Two datasets of 10 million records each have 100 trillion potential links between them meaning scalability is imperative. The result from Splink has similar accuracy to some of the best alternatives, but faster, at greater scale, and with more flexibility. The package is publicly available. There are online demonstrations of the various customisation options available, as well as information on how to run the code and apply it to any dataset.

The data linking methodology for external data shares is agreed between the parties, based on the common identifying information available. For example, linking for the MoJ-DfE share used a deterministic approach, developing matching rules using common variables between the different sources. Matching rules included combinations of at least an exact match on three of the five variables available as well as applying ‘fuzzy matching’ techniques to names.

A separate data linkage is carried out when datasets are shared with the SAIL Databank. This is carried out by a Trusted Third Party and allows MoJ data to be matched to other datasets held by SAIL.

Assigning a match

Data First has deduplicated records within each MoJ dataset using the Splink package, calculating match scores for millions of pairs of record comparisons. Record pairs with match scores above a specific threshold have been designated a match, and are grouped together with a common unique defendant ID. This new derived column is added to datasets used by researchers.

Links between records across the MoJ datasets are then made and shared in a separate linking dataset that acts as a lookup between the datasets. This linking dataset is extended as additional data are incorporated.

Without a labelled training dataset (where it is known which records refer to the same individual) assessing the accuracy of data linking is challenging. To set an optimal threshold match score for the magistrates’ courts data, tens of thousands of record pairs have been clerically scored to simulate this ideal labelled data. Providing detail on

matching probabilities is not straightforward as these are calculated pairwise and would require a full dataset of 160 million potential matches to analyse. Information on the strength of the match will therefore not be made available in standard datasets.

As mentioned above, where a pair of records matches, they are assigned to a group, and any further matches to either of them is assigned to the same group to represent a single defendant. No attempt is made to define a canonical record of personal information (determine which is the correct/current value of name, address etc.) but just to assign a common unique defendant ID. This is the key variable for researchers to use in addressing questions about patterns of activity for justice system users (for example, repeat defendants).

Limitations of data linking

Given the limitations of the personal identifying information in source data it is sometimes not possible to know with certainty whether two similar records relate to the same person. The choice of threshold match score will always represent a trade-off between the risks of false positives (linking records which belong to two different people) and false negatives (not linking records which do belong to the same person); each has different implications for research.

The probability threshold used should be suitable for most research and statistical purposes, for example, providing sensible estimates of the frequency of repeat interactions (individuals returning to court in a given time frame), and insights into shared characteristics of individuals with similar patterns. However, it is expected that a small proportion of false links will be included, where records belonging to two or more people are erroneously attributed to one person, and that not all genuine links will have been made due to the matching probabilities previously mentioned.

The quality, consistency, and uniqueness of source data about individuals affect data linking accuracy. For example, it is much more difficult to determine that records belong to one person if they have used different names and moved address often, while names that appear less frequently can be grouped together more confidently. Researchers should be aware that accuracy in data linking for groups with different characteristics (such as socio-economic status or ethnicity) could differ because of these factors.

The proportions of false and missed links varies between datasets and is dependent on the completeness of demographic variables used during the linking process. In the case of family courts for example, missingness of key fields has led to non-randomness in linking taking place, and linkage being dependent, at least in part, on an individual's role in a case. This will need to be considered as part of designing, reviewing, refining, and approving research projects. Researchers are welcome to contact us to discuss project ideas at their earliest convenience so we can advise on viability.

How is the data protected?

The Data First programme has been developed within the framework established under the Digital Economy Act (DEA) (2017) which enables government to prepare

administrative data for the purposes of research, and to provide de-identified versions of those data to researchers and projects accredited by the UK Statistics Authority (UKSA).

Data First is not sharing data directly with individual researchers but making pre-defined extracts securely available through our partnerships with the ONS Secure Research Service and SAIL Databank and using the Five Safes Framework to ensure the safety and security of its stored data. This is a set of principles adopted by a range of secure labs to provide complete assurance for data owners.

The Five Safes are:

- Safe People – trained and accredited researchers are trusted to use data appropriately. To access the data individuals must complete training and become accredited researchers under the DEA.

- Safe Projects – data are only used for valuable, ethical research that delivers clear public benefits. To access data the research project must have been approved by both the data owners (data access panels at MoJ and its agencies) and the Research Accreditation Panel (RAP). In order for research projects to be approved they must comply with the Research Code of Practice and Accreditation Criteria which was approved by the UK Parliament in July 2018.

- Safe Settings – access to data is only possible using secure technology systems through our partners at the ONS Secure Research Service and the SAIL Databank. These are Accredited Processing Environments under the DEA which meet rigorous standards of security and data capability controls.

- 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. Personal identifiers used in the linking are not made available for analysis and replacement values are generated for internal system IDs to prevent direct linkage back to the raw data source. Some special category data is being shared because of its value to research. It is highly unlikely that a researcher will be able to identify individuals from these fields alone.

Data processing within Data First is compliant with all applicable data protection legislation, including the General Data Protection Regulation and Data Protection Act 2018, and a suitable legal gateway is required for all external data linkage. The MoJ-DfE data share does not rely on powers in the DEA.

Datasets available

Table *1* summarises the datasets currently available through Data First, including the source(s) from which they are derived, time period covered, key variables, potential research areas, which organisation(s) hold the data and what approvals are required to

access it. More detailed information on dataset variables can be found within the relevant Data Catalogues, available on GOV.UK.

Table 1 - Summary of information on available datasets

Name	Source	Dates covered (latest version)	Key variables	Potential research areas	Approval required	Data held by	Notes / Additional information
MoJ Data First cross-justice system linking dataset	Multiple sources. Created from linkage processes using personal information not shared from individual datasets below.	January 2011 – March 2023 (depending on record source)	• Cross-justice system linking table: Unique person ID connecting person records between other Data First datasets. • Criminal courts case-level linking table: connecting case records between magistrates' courts and Crown Court datasets.	• Cross-cutting questions and user journeys. • Research focusing on groups interacting with different justice services.	Depends on data requested alongside	ONS SRS and SAIL Databank	To be used alongside other data to allow joining between multiple Data First datasets. Does not contain information about people or cases. Acts as a lookup between individual datasets below.
MoJ Data First magistrates' courts defendant dataset	Libra (magistrates' courts management information system)	Case disposals (completions) from January 2011 – March 2023	• Unique defendant ID • Defendant characteristics • Case and Sentencing Outcomes • Offence categorisation • Cases and hearings • Key outcomes • Case timings • Most serious offence flag (in case)	• Nature and extent of repeat users of the magistrates' courts • Who are our repeat users? • Sentencing changes with repeat use of the magistrates' courts • At what stage in a case do people plead guilty? • Differences in experience and outcome by characteristics such as the	HMCTS Data Access Panel (DAP)	ONS SRS and SAIL Databank	Part of cross-justice system linkage. This dataset contains the following tables: • hocas_flatfile – one row per defendant per case, based on the most serious offence in a case • hocas_all_offence – one record per offence dealt with by the courts, covering all offences

					defendant's age and ethnicity			
MoJ Data First Crown Court defendant dataset	Xhibit (Crown Court management information system)	Case disposals (completions) from January 2013 – March 2023	• Defendant ID • Defendant characteristics • Warrants • Remands • Key outcomes • Most serious disposal flag • Pleas • Bail	• How do previous convictions affect sentencing in the Crown Court? • Impact of earlier pleas on sentencing outcomes in the Crown Court • Associations between defendant characteristics and being sentenced to prison in the Crown Court • Key drivers of delays in the Crown Court • Repeat users of the Crown Court, their characteristics, how often do they return, how long does it take for repeat users to return?	DAP	ONS SRS and SAIL Databank	Part of cross-justice system linkage. This dataset contains the following tables: • xhibit_flatfile – one row per defendant per case, based on the most serious offence in a case • xhibit_all_off_disp – one record for each offence and disposal within a Crown Court case, covering all offences
MoJ Data First prisoner custodial journey level dataset	P-NOMIS (Prison National Offender Management Information System)	Custody period spanning January 2011 - March 2023 (custodial sentences started since 2011 expected to be complete,	• Unique offender ID • Characteristics of offender • Main offence • Sentence • Release information (if applicable)	• Characteristic profile of repeat occupiers of the prison system • Are certain release types more likely to deter an offender from re-entering the system? • Association of characteristics	Data Access Group (DAG) and Data Access Governance Board (DAGB)	ONS SRS and SAIL Databank	Part of cross-justice system linkage. This dataset contains the following tables: • nomis_flatfile – one record per prisoner per custodial journey • nomis_movements – one record per prisoner per movement between prisons or in / out of prison

		but earlier sentences are included)			such as ethnicity and gender with variation in custodial reconviction			• nomis_safety_in_custody – one record per prisoner per incident where assault or self-harm questionnaires are completed
MoJ Data First probation dataset	nDelius (National Delius)	Referral dates (start of probation supervision) from January 2014 – March 2023	• Offender characteristics • Community order requirements • Licence conditions • Post sentence supervision requirements • Recalls • Terminations and breaches	• Factors affecting the likelihood of different groups receiving different sentences, variations in sentencing recommendations by the availability of different options • Enablers and barriers to effective sentences, including community-based, alternative and short custodial sentences • Changes in non-custodial sentencing • Effectiveness of electronic tagging and monitoring in protecting the public from harm; are there specific groups of individuals for who electronic tagging and monitoring is more effective? • Addressing anti-social, violent, and criminal behaviour linked to alcohol	DAG and DAGB	ONS SRS and SAIL Databank	Part of cross-justice system linkage. This dataset contains the following tables: • delius_flatfile – one row per offender event • delius_rqmnt – one row per requirement • delius_lic_condition – one row per license condition • delius_psr – one row per court report or proposed requirements • delius_pss – one row per post sentence supervision requirement

	and drug use beyond traditional criminal sentencing • Impact of home detention curfew, in advance of custodial sentence completion, on individual outcomes and risk to public protection; how can home detention curfew be improved? • Do licence period conditions and durations affect the potential for recalls, and the impacts on individual outcomes and risks to public? • Can short periods in custody be made more effective at reducing reoffending? • Effectiveness of longer custodial sentences on crime? • Effectiveness of rehabilitation activity requirements

					(RARs) and possible improvements			
MoJ Data First family court dataset	FamilyMan (family court management information system)	Cases active from January 2011 – March 2023	• Unique ID for party • Divorce, private law and Family Law Act cases • Public law and adoption cases • Applications and orders made • Marriage and divorce characteristics • Party characteristics • Case events, timelines and number of hearings • Legal representation	• Nature and extent of repeat use of the family court • Who are repeat users, what are their demographics, and do the same parties return to the court in different contexts? • Public and private law overlap; characteristics of these users, pathways through the court and the outcomes for these groups	DAP	ONS SRS and SAIL Databank	Part of cross-justice system linkage PLEASE NOTE: Due to limitations in the linking of this dataset (as a result of missing data items available to the linking process), overarching or macro-level research in particular is likely to be subject to systematic bias resulting from false-negatives (missed connections). We ask that any researcher planning to use this dataset in their research, contact us prior to submitting an application, to discuss project ideas at their earliest convenience and so we can advise on project viability. Linked to Cafcass and Census 2021 England and Wales data in SAIL This dataset contains the following tables: • family_cases – one row per case • family_events – one row per event within a case • family_people – one row per party for each case they are involved in
MoJ Data First civil court dataset	CaseMan (county court case management information system) and PCOL (Possession Claim online service)	Cases started from January 2012 - January 2022	• Unique ID for parties who are individual court users • Mortgage and landlord possessions • Money claims, damages, bankruptcy	• Nature and extent of repeat use of civil courts • Who are repeat users, and do the same parties return to the court in different contexts?	DAP	ONS SRS	Part of cross-justice system linkage. Some cases will appear in both CaseMan and PCOL tables and can be linked, for example where a possession case begun on PCOL progresses to hold hearings in a local county court. This dataset contains the following tables (for both CaseMan and PCOL):

			• Case events and timings • Hearings and venues • Judgements awarded • Warrants and enforcement • Legal representation	• How do different types of civil case progress and reach resolution? • How often are different types of civil cases defended and parties represented? • How do levels of judgement awarded differ?			• cases / claims – one record per case • case_events / claim_events – one row per event • hearings – one row per hearing (there may be several hearings in a case) • judgments – one record per judgment made • warrants – one record per warrant issued • parties – one row per party for each role in each case they are involved in
MoJ Offender Assessment dataset	OASys	1 January 2011 – 31 December 2023 (contains assessment data up to 31 December 2023, but only records up to 31 March 2023 can be linked to other data sources via the cross-justice system linking dataset)	• Unique ID • Offender Characteristics • Offending information • Risk Scores • Analysis of Offence(s) • Information on accommodation needs • Education, training & employability needs • Financial Management & Income • Relationships • Lifestyle & Associates • Drug Misuse	• How can we better understand how problems interact and reinforce each other and how people move through different jurisdictions across the system as they attempt to resolve them? • How do individual's needs change over time within the criminal justice system? • What characterises individuals that stop offending? • What are the risks and needs of	DAG and DAGB	ONS SRS (Work ongoing to make available via the SAIL Databank)	Part of cross-justice system linkage. This dataset contains the following tables: • assessments_metadata – one row per assessment and sub-assessment • oasys_assessments_details_bcs – one row per basic custody screening • oasys_assessments_details_core_risk – one row per layer 1, layer 2, layer 3, oasys 1 and oasys 2 assessment • oasys_assessments_details_isp – one row per initial sentence plan meeting record • oasys_assessments_details_rm2000– one row per risk matrix 2000 sub-assessment which can be linked back to the corresponding parent assessment

			• Alcohol Misuse • Emotional Well-being • Thinking & Behaviour • Attitudes • Health and other considerations • Information on Sentence Planning	different offender groups? • How do risks and needs change over time and affect their likelihood of reconviction? • How do the characteristics of specific offender cohorts such as sex offenders or perpetrators of domestic violence differ from other offenders?			• oasys_assessments_details_rmp – one row per risk management plan • oasys_assessments_details_rosh – one row per risk of serious harm screening (sections R1-R5) • oasys_assessments_details_rosh full – one row per risk of serious harm full analysis (sections R6-R9) • oasys_assessments_details_rosh sum – one row per risk of serious harm summary (section R10) • oasys_assessments_details_rsp – one row per review of a sentence plan • oasys_assessments_details_saq – one row per self-assessment questionnaire • oasys_assessments_details_sara – one row per spousal assault risk assessment sub-assessment which can be linked back to the corresponding parent assessment • oasys_assessments_details_skill schecker – one row per record of the skills checker tool • oasys_assessments_details_tr_bcs – one row per basic custody screening following the transform rehabilitation reforms of December 2014
MoJ-DfE data share	PNC (Police National Computer) and NPD (National Pupil Database)	2000 – 2017; offenders born after 31st August 1985	• Unique ID • Academic achievement • Pupil absence • Pupil exclusion • Criminal history • Court appearances • Time in prison	• Association between particular interactions with the education system and offending; is one of these factors typically the driver?	DAG and DAGB **plus DfE approval process**	ONS SRS	Separate data linkage. Access conditions and processes may differ from other datasets. Can only be accessed through ONS SRS. Data catalogue available on request.

		• Relationships between educational and criminal justice outcomes, and impact by demographic factors • Are interventions to prevent Serious Violence effective (through use to generate control groups)?

Section 3: Applying for data access

Accessing Data First data

Access is available to justice data, and external linked data where agreed, through two Trusted Research Environments (TREs): the ONS Secure Research Service (SRS) and SAIL Databank, which are both Accredited Processors under the Digital Economy Act (2017) and have substantial expertise in data management, metadata and the checking of outputs.

The datasets are accessed through these secure platforms and only approved outputs are removed from the systems.

Both the researcher and their specific research project must be approved and accredited before access is granted to the data.

From a researcher's perspective there are three steps needed to gain approval to access the data:

1) Become an accredited researcher.

2) Apply for access to the specific data required for a research project from relevant data owners using the MoJ and HMCTS 'Secure access to data' combined form for access to data held in the ONS SRS, or complete the scoping form and Information Governance Review Panel (IGRP) form for data held by SAIL Databank.

3) Complete the Research Project Application form and receive final approval from UKSA RAP (not required for MoJ-DfE data share).

Becoming an accredited researcher

Applicants requesting access to Data First datasets, either through the ONS SRS or SAIL Databank, will need to be an accredited researcher. It is recommended that you begin this process as soon as possible and in parallel to any research application.

To access datasets from the ONS SRS, applicants will need to be accredited under the ONS accredited researcher scheme. To become an accredited researcher, you must download the relevant forms listed on the UK Statistics Authority website and submit them electronically to Research.Support@ons.gov.uk to begin your journey.

ONS' Research Services and Data Access (RSDA) team will support all researchers throughout the journey (see Figure 4). The RSDA team creates and looks after all the

training requirements for those looking to become accredited researchers. The RSDA team advises and supports researchers who have passed the course and achieved their accredited status on the drafting of research proposals, providing them with advice from ONS experts.

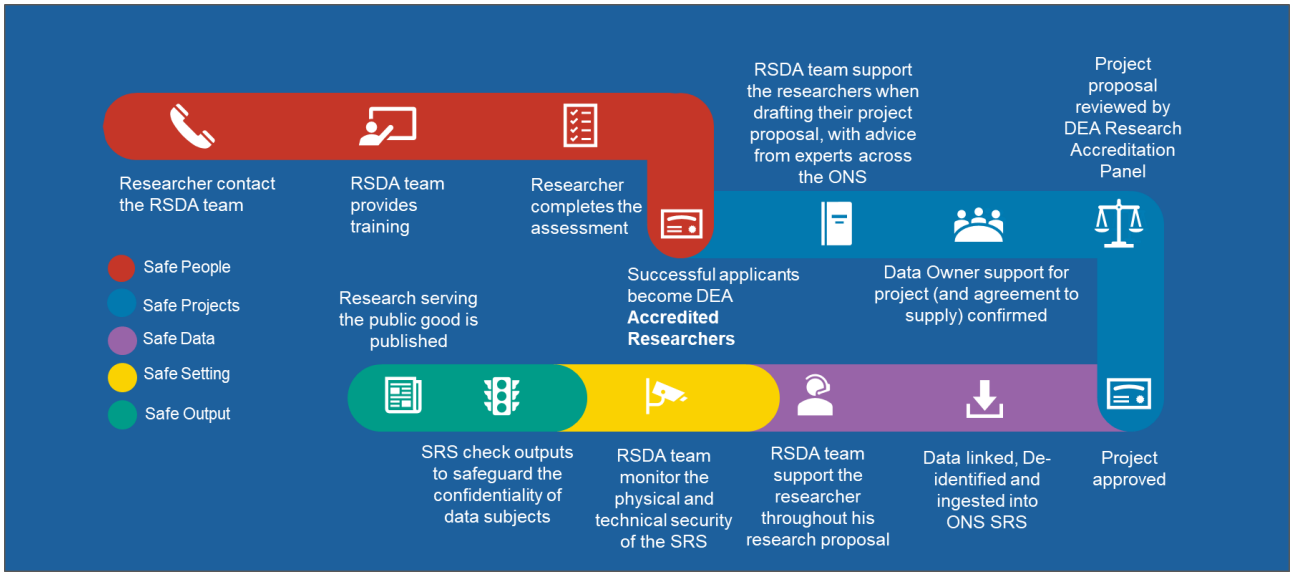

Figure 4 – A walkthrough of a researcher's accreditation and project development journey when applying to ONS SRS

Applicants wishing to access data via SAIL Databank can either complete the ONS accredited researcher scheme, a UK Research and Innovation (UKRI) Medical Research Council (MRC) e-learning module (Research, GDPR and Confidentiality – what you really need to know) or other equivalent training. Users must also provide evidence of a research career, via a CV, and belong to a recognised institution. Further information on safe researcher training can be found on the relevant section of the SAIL Databank website.

Data First research governance processes

Approval from the relevant data owner must be gained to access data for a research project. The majority of Data First datasets are available through both the ONS Secure Research Service (SRS) and SAIL Databank.

There are different approval processes for accessing Data First datasets via the ONS SRS or SAIL Databank, and for HMPPS and HMCTS data. The governance processes for seeking approval are outlined in Figure *5*.

Applications seeking access to courts data require approval from the HMCTS Data Access Panel (DAP), while access to HMPPS data is given by the MoJ's Data Access Governance Board (DAGB), following preliminary assessment by the Data Access Group (DAG). Applications for both HMPPS and HMCTS data will be assessed by DAG, DAGB and DAP.

Approval from other government departments (OGDs) must also be obtained for any external linked data (for example, the MoJ-DfE data share).

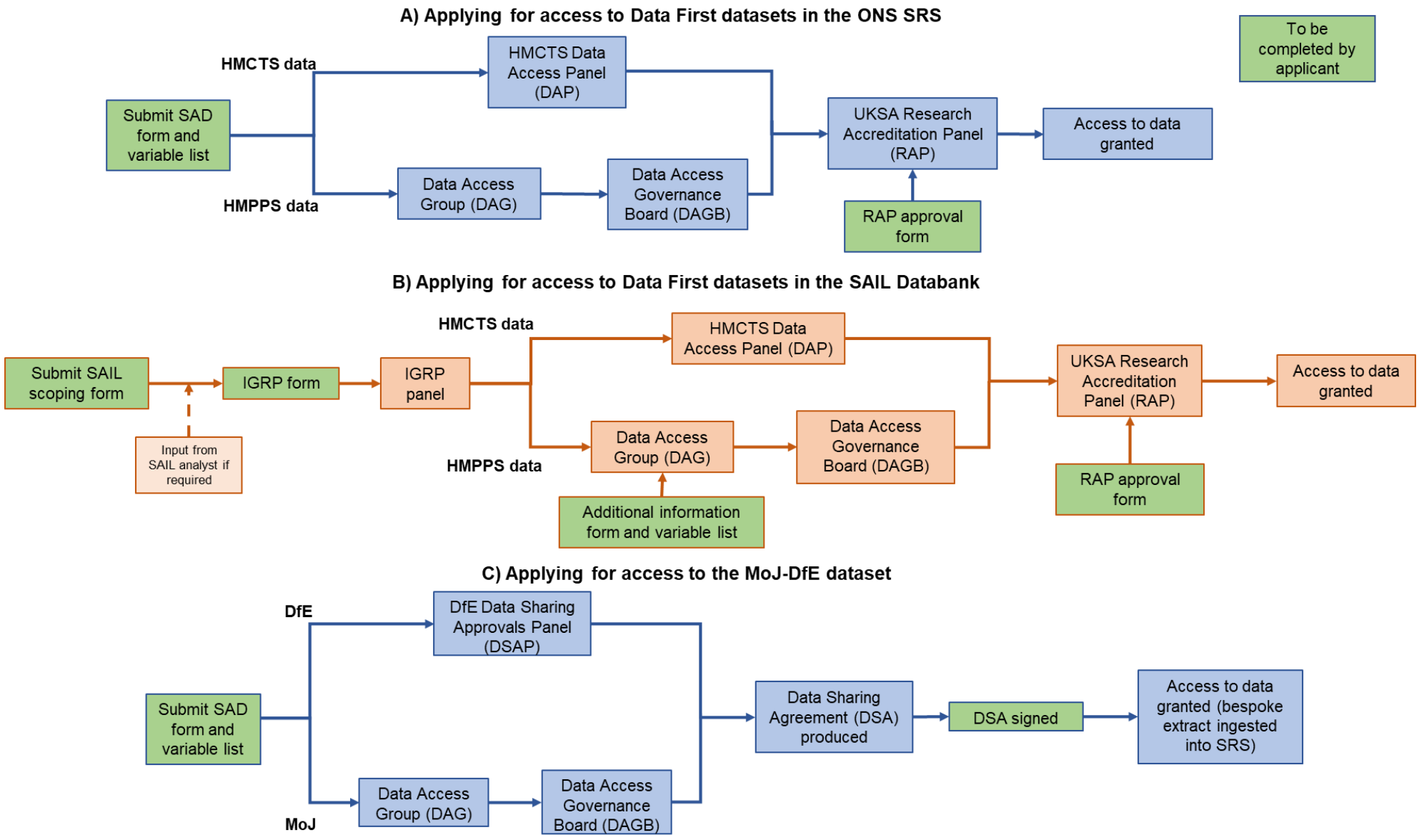

Figure 5 - The process for gaining access to Data First datasets through A) ONS SRS, B) SAIL Databank and C) for accessing the MoJ-DfE dataset

A single Secure Access to Data (SAD) application form combining requests to both the DAGB and the DAP is available on GOV.UK, and combines requests for access under Data First and other MoJ and HMCTS data. A separate guidance document and the "available data and their modes of access" ODS file directs researchers to which sections of the application form to complete, the data available and provides contact information detailing where the form should be sent to.

Applicants to SAIL Databank should first complete a scoping form outlining their proposed research and will receive input from a SAIL analyst to complete the Information Governance Review Panel (IGRP) form. An additional supplementary form needs to be completed for access to Data First datasets which is approved by the relevant data owners and is an independent step from the IGRP approval process.

For all project applications an accompanying file must be provided with application forms that lists the variables required from each dataset intended to be used and a justification for each variable. The justification can be kept simple and concise but should be clear in what the variable is being used for and why. Applications will be reviewed by analysts and data experts within MoJ and its agencies who will judge whether the research proposal meets the required criteria. The panel considers a number of criteria:

- Is the research proposal ethical (including its potential impact on data subjects)?
- Is the data necessary to address the research questions?
- Have data protection concerns been adequately addressed?
- What is the overall benefit of the research?

The panel's recommendations will be scrutinised by members of the relevant approval panels who will decide if the researcher will be granted access to the data requested.

The Data First programme will work with an internal ethics advisory group at MoJ, external ethicists and the Academic Advisory Group (AAG) to develop our guidelines on ethical use of the data. Additionally, the Data First team will consult with the Data First User Representation Panel, which brings together organisations that represent justice system users, to ensure public acceptability of the work.

After passing through the relevant panel(s) within MoJ and SAIL Databank, and/or OGD processes as applicable, final approval for the project must be obtained from UKSA RAP. RAP ensure the independence, consistency and transparency of the process for granting accredited researchers access to de-identified data, and apply criteria of their own to assess an application:

- Is there a public benefit?

- Is there demonstrable analytical merit?
- Is the project feasible?
- Are any relevant privacy implications sufficiently mitigated?
- Has the project successfully completed a formal ethical review?

Documentation and guidance on the RAP approval process can be found through the UKSA website. Applications should be submitted via email, and can be started in parallel with the MoJ and/or OGD approval process, but will not be reviewed until approval has been obtained from the relevant departmental bodies.

A brief summary of successful applications to access the data can be found through the Data First application webpage.

Choosing where to apply

The MoJ-DfE data share can only be accessed through the ONS SRS.

Data First cross-justice system datasets are being made available through both the ONS SRS and SAIL Databank. Applicants should consider a number of factors when choosing where to apply for access.

Considerations may include:

- Existing arrangements your university / institution has that may make access to the data more straightforward (for example an Assured Organisational Connectivity agreement)
- Whether you need to be in a particular physical location in order to access the data (for example via a Safe Room or SafePod network)
- Specific expertise of the organisation you apply to
- Available software for analysis
- Any other relevant data you wish to use; for example, within SAIL, Data First family court data is already linked to Cafcass and Census 2021 data. Data First cross-justice system datasets can potentially be linked to other datasets such as a range of Welsh health and population data. SAIL can advise on availability and linkage of non-MoJ sources.

Further detail and advice can be provided as part of the application process.

How do I access my data?

ONS SRS

Once researchers and their projects are accredited and approved, projects using the SRS will have a project space created. Datasets requested for projects will be mapped to the project space. Researchers may also send data to the ONS Statistical Support Team to be added to the project space, which they will receive guidance on if they choose to do so.

Researchers named on projects will then be provided with their account details and instructions on how to access the SRS. Access to the SRS is through a safe setting. Safe settings may be in safe rooms on ONS sites, in safe rooms on other certified sites, or through an organisation which has an Assured Organisational Connectivity Agreement with ONS, and which maintains a current certification.

For more information on this process contact ONS at research.support@ons.gov.uk

ONS SRS - Tools for analysis

Research is conducted in the SRS environment using software that has been tested and installed by the SRS operations and security team. The SRS makes every effort to provide software that is as up to date as possible. Table *2* is a list of the software that is currently available to use for researchers:

Table 2 - Software available for researchers through ONS SRS

Software
Adobe Acrobat Reader DC
Anaconda 4.4
Python 3.7.6
ArcGIS 10.4.1
ArcMap 10.4.1
Jupyter Notebook
Microsoft Office Professional Plus 2021
ML-Win (3.04)
Notepad++
Qtconsole
R for Windows (3.6.1, 4.0.2, 4.3.0)
R Studio (2022.07.2)
SAS v9.3
SPSS v29
Spyder 4
STATA (16MP, 17SE)
Winzip
MPlus 8 Demo
SQL Server Management Studio
Java Development Kit
JAGS

Researchers can also request that code they have written is ingested into their project space. ONS are currently unable to ingest packages from open-source code repositories such as CRAN or GitHub.

SAIL Databank
Following researcher accreditation and approval, users will be able to access the SAIL Gateway, which has been designed to provide a familiar Windows environment with an array of toolsets and applications.

Researchers can access their project space through a remote desktop connection, or alternatively through the Safe Pod Network (SPN).

Users may seek permission to import non-data files such as syntax scripts and reference documents if necessary.

SAIL Databank - Tools for analysis
SAIL make a range of tools available as standard, including SQL querying tools and other commonly used applications such as MS Office, SPSS and R. Other applications, such as STATA, are available for a small licensing fee, and users can also request the addition of other applications for which they hold a license.

A full list of all software, including versions, is available from SAIL on request.

Further Information

General contact and enquiries	datafirst@justice.gov.uk
OGD data shares	datalinkingteam@justice.gov.uk
MoJ Data First webpage	Ministry of Justice: Data First – GOV.UK (www.gov.uk)
MoJ and HMCTS Data Access form and guidance	https://www.gov.uk/government/publications/moj-data-first-application-form-for-secure-access-to-data
ONS accreditation	https://uksa.statisticsauthority.gov.uk/digitaleconomyact-research-statistics/better-useofdata-for-research-information-for-researchers/
ONS Research Services and Data Access (RSDA) team	research.support@ons.gov.uk
Data First on ADR UK	https://www.adruk.org/our-work/browse-all-projects/data-first-harnessing-the-potential-of-linked-administrative-data-for-the-justice-system-169/
MoJ Areas of Research Interest	https://www.gov.uk/government/publications/ministry-of-justice-areas-of-research-interest-2020
Splink data linking package	https://github.com/moj-analytical-services/splink
Splink webpage	Splink (moj-analytical-services.github.io)
SAIL Databank	https://saildatabank.com/
SAIL Databank contact form	https://saildatabank.com/contact/
UKSA Research Accreditation Panel	Research Accreditation Panel – UK Statistics Authority