Ministry of Justice

Data First:

An Introductory User Guide

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

Version 7.0, February 2022

ADR UK

Contents

What is this guide for?

This user guide is intended to support researchers in seeking access to data that has been shared under the Data First project. It provides an overview of the project and forthcoming data-shares, including information on data content, quality and security.

The guide also provides detailed guidance on the processes required to access shared data. This include the process of becoming an accredited researcher as well as the Data First research governance processes to seek approval for access for specific research projects.

Information on how to access the Office for National Statistics (ONS) Secure Research Server (SRS) and the resources for analysis that will be made available for researchers are also included.

What is the Data First project?

Project overview

Data First is an ambitious data-linking 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) (see Figure 1).

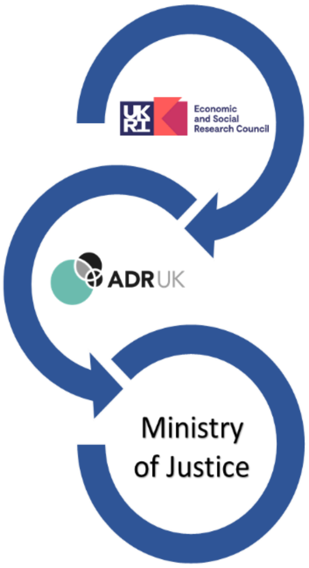


Figure 1: The link between different organisations in the setup of Data First

Data First aims to unlock the potential of the wealth of data already created by MoJ, by linking administrative datasets from across the justice system and beyond, and enabling accredited researchers, across government and academia, to access anonymised, research-ready datasets ethically and responsibly. The project will also enhance the linking of justice data with other government departments (OGDs).

By working in partnership with academics, other government departments, and wider justice sector organisations to facilitate research in the justice space, we will create a sustainable body of knowledge on justice system users and their needs, pathways and outcomes across public services. This will provide evidence for the development of government policies and progress the tackling of 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 will enable some of the critical evidence gaps outlined in the ARI to be explored in collaboration with our partners. In doing so, the aim is to strengthen the strategic research capabilities across government and academia and reinforce the impact of evidence at all stages of policy development and evaluation.

What is the potential of this newly linked data?

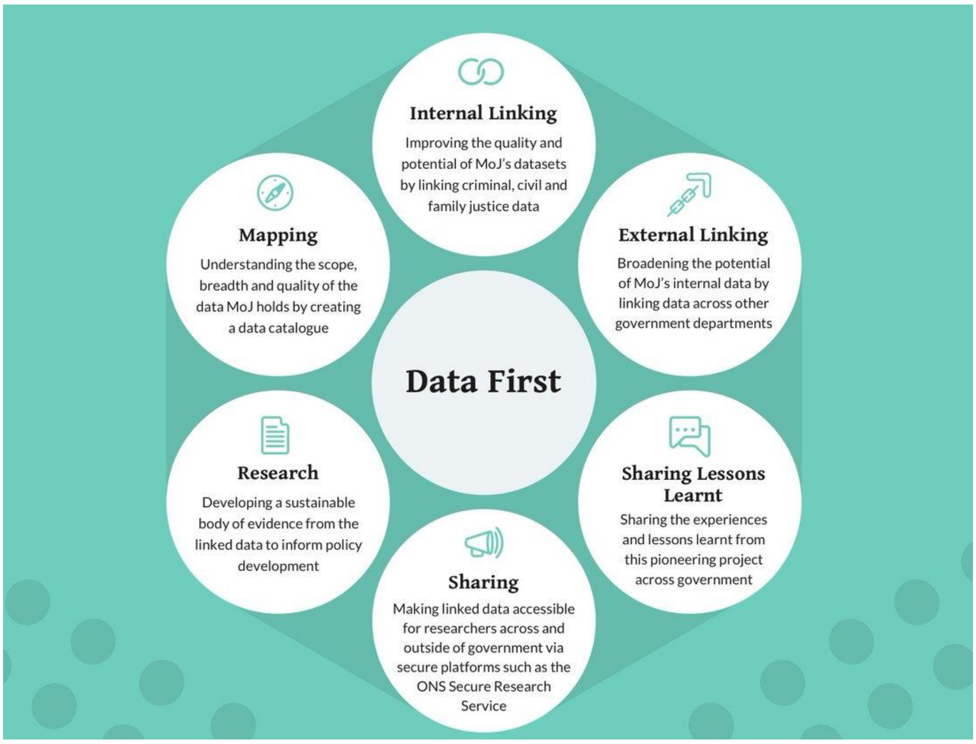
By linking the civil, family and criminal justice administrative datasets, we can build a picture of justice system users and interactions over time across the courts, prison and probation services. Understanding these characteristics, patterns of frequent use, and common transitions between different 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 from these newly linked datasets. Data First is designed to facilitate links with our research and academic partners; by working in collaboration we will identify priority areas for analysis to make best use of the data. This could provide insight on, for example, repeat users of the justice system and its services, helping us develop a better understanding of sentencing outcomes and improving user experiences.

Linking MoJ data with that of OGDs will enhance our understanding of how justice system users interact with other public services, and their needs, pathways and outcomes across a range of events.

What does Data First involve?

Data First is comprised of several different elements to ensure that the project reaches its full potential (see Figure 2).



Data First

Internal Linking

Improving the quality and potential of MoJ's datasets by linking criminal, civil and family justice data

External Linking

Broadening the potential of MoJ's internal data by linking data across other government departments

Sharing Lessons Learnt

Sharing the experiences and lessons learnt from this pioneering project across government

Sharing

Making linked data accessible for researchers across and outside of government via secure platforms such as the ONS Secure Research Service

Research

Developing a sustainable body of evidence from the linked data to inform policy development

Mapping

Understanding the scope, breadth and quality of the data MoJ holds by creating a data catalogue

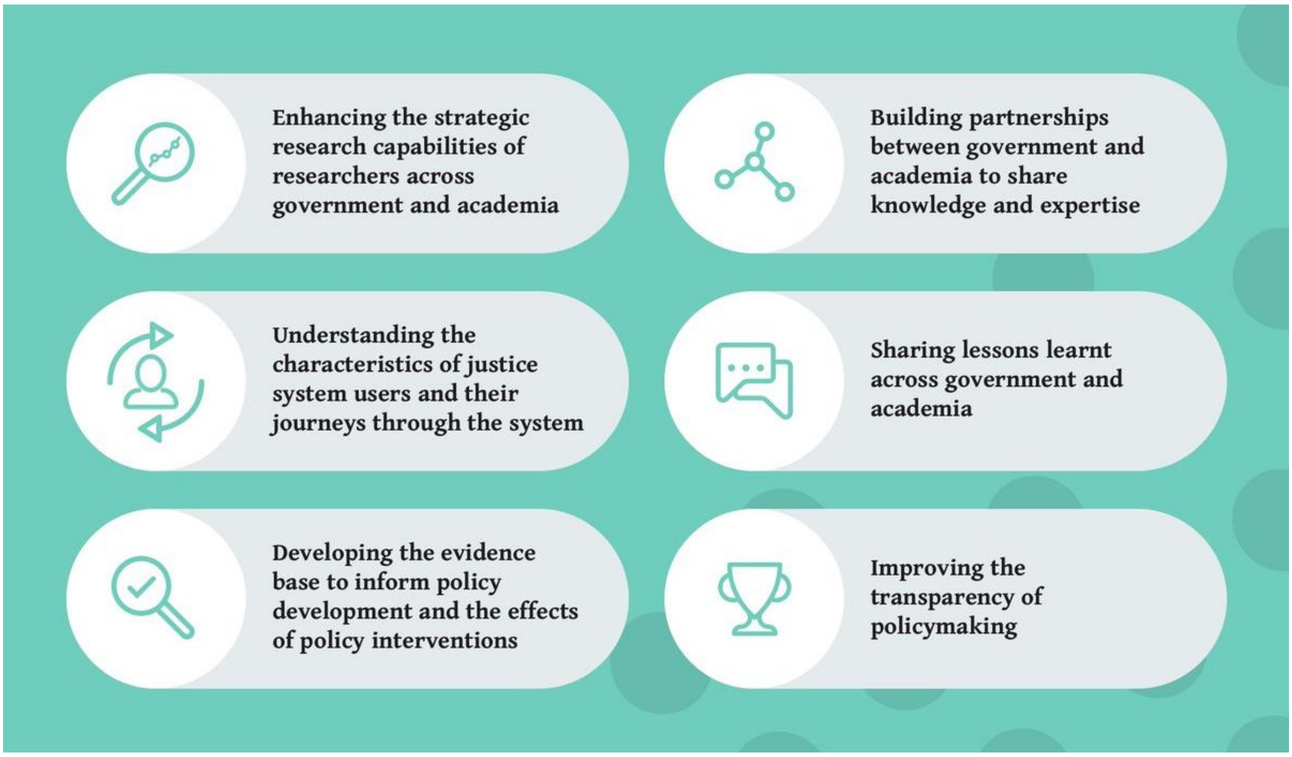
Figure 2: Representation of the different workstreams within Data First

There are four internal teams leading on different workstreams of the project within MoJ:

- **Internal data linking** – a team of data scientists and data engineers leading on the development of a robust, automated linking pipeline between criminal, family and civil justice datasets.
- **External data linking** – a team of statisticians and operational researchers leading on establishing data-shares with external partners and linking justice data with OGDs.
- **Data mapping and strategy** – a team of social researchers and statisticians leading on mapping data held across MoJ with a view to developing an externally-shareable list of research-ready datasets.
- **Research, academic engagement and communications** – a team of social researchers and statisticians who are facilitating the link between Data First and the research and academic community, 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 (outlined in Figure 3).



Enhancing the strategic research capabilities of researchers across government and academia

Building partnerships between government and academia to share knowledge and expertise

Understanding the characteristics of justice system users and their journeys through the system

Sharing lessons learnt across government and academia

Developing the evidence base to inform policy development and the effects of policy interventions

Improving the transparency of policymaking

Figure 3: Benefits of Data First

- **Research** – the project looks to enhance the strategic research capabilities of researchers across both government and academia.
- **Better understanding of justice system users** – the project's linkage of data can help to provide 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 project will allow MoJ to share the lessons they learn whilst establishing and delivering Data First across government and academia to share the learning with our stakeholders.
- **Improving the transparency of policy-making** – research published under the Data First project will further enhance the openness of the use of evidence in policymaking.

What is included in the data-shares?

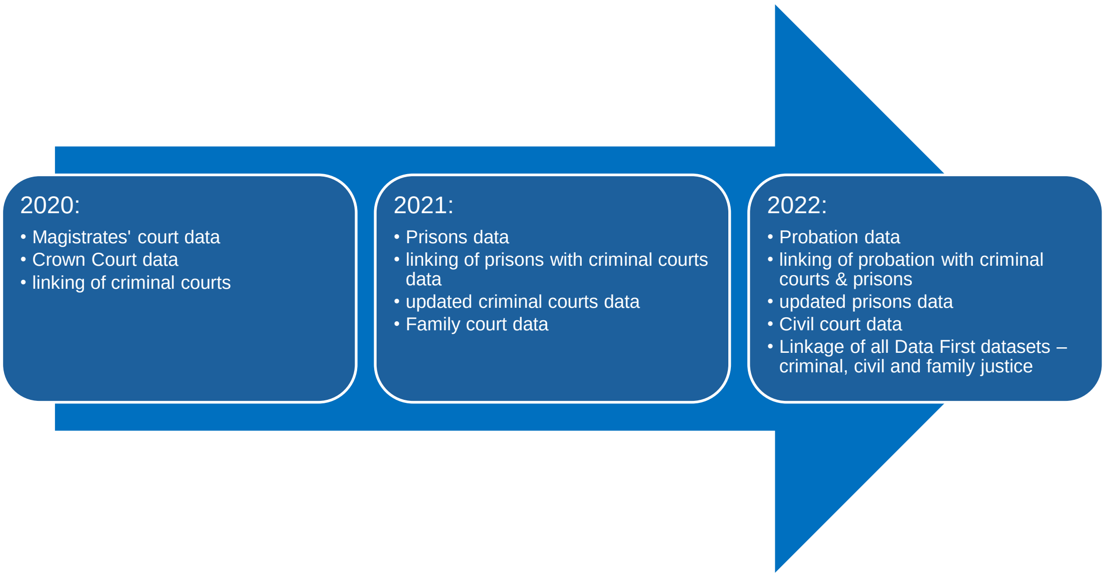
How do these shares differ from the original datasets?

The Data First project includes work to structure 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 we do not anticipate material differences in trends or conclusions, researchers should be aware that analysis carried out using Data First datasets may not be exactly comparable to other published statistics or research.

Provisional data-share timeline

Data First is currently a three-year project running until 2022, with plans to share different criminal, civil and family justice datasets throughout its duration.

Over the first two and a half years the project has made datasets available containing data from the magistrates’ and Crown Court, prisons, and the probation service and completed linking of people involved across these stages of the criminal justice system. Data from the family courts has also been released identifying where people are involved as parties in multiple family court cases. The development of a civil justice dataset is underway, with a linkage across all criminal, civil and family justice datasets anticipated by the end of 2022 (see Figure 4).



2020:
- Magistrates' court data
- Crown Court data
- linking of criminal courts

2021:
- Prisons data
- linking of prisons with criminal courts data
- updated criminal courts data
- Family court data

2022:
- Probation data
- linking of probation with criminal courts & prisons
- updated prisons data
- Civil court data
- Linkage of all Data First datasets – criminal, civil and family justice

Figure 4: Provisional timeline of potential Data First internal shares

External data-shares involve the linking of MoJ data with data from OGDs. Data First is supporting academics to access and carry out research using the MoJ-Department for Education (DfE) linked dataset which was made available in May 2020. The development of this share (linking MoJ's Police National Computer and DfE's National Pupil Database) pre-dates the Data First project and is not funded by ADR UK.

The project is planning to agree further external shares during 2022.

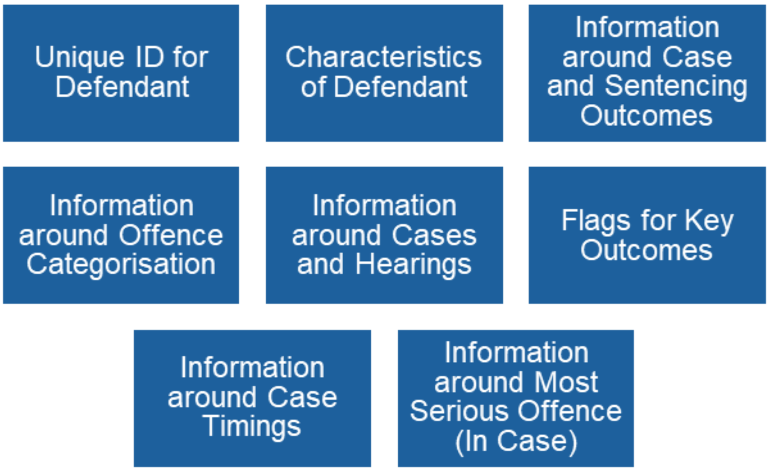
These data-share timelines are provisional and subject to change depending on factors such as the feasibility of data linkage or finalisation of data sharing agreements between parties.

Further information on the timing and content of forthcoming data-shares will be provided in future updates of this guide. Details of external data shares will be announced once an agreement has been reached between the parties.

MoJ Data First magistrates' court defendant case level dataset

What does this share consist of?

The magistrates' court defendant case level dataset provides data on defendant appearances in the magistrates' court between January 2011 and December 2020 and is extracted from the magistrates' court management information system (Libra).



Unique ID for Defendant

Characteristics of Defendant

Information around Case and Sentencing Outcomes

Information around Offence Categorisation

Information around Cases and Hearings

Flags for Key Outcomes

Information around Case Timings

Information around Most Serious Offence (In Case)

Figure 5: Descriptions of variables in the magistrates' court dataset

The dataset has been deduplicated, which means multiple instances of the same person appearing before the magistrates' court over the period are identified and assigned a single unique identifier.

This enables researchers to establish if the same user has entered the criminal courts more than once, and the frequency, purpose and outcomes of those appearances.

The magistrates' court dataset includes approximately 60 variables, including the dates of hearings, type of offences, arrests and the initiation date of proceedings (see Figure 6). This data explains how proceedings were dealt with (for example, whether the defendant chose trial by jury) and provides information on the final plea of the defendant.

Other data include details on bail, remand, and whether the defendant was committed to the Crown Court.

What is the potential of this data?
Sharing this deduplicated data with accredited researchers will help develop the evidence base on the magistrates' court defendants and their outcomes.

Examples of research questions that this might be able to answer:

- What is the nature and extent of repeat users of the magistrates' court?
- Who are our repeat users?
- How does sentencing change with repeat use of the magistrates' court?
- At what stage in a case do people plead guilty?
- How do experiences and outcomes in the magistrates' court differ by characteristics such as the defendant's age and ethnicity?

As the Data First project continues, magistrates' court data will be linked to other datasets in order to address other research questions. For example, case progression to the Crown Court and prison and probation services.

MoJ Data First Crown Court defendant case level dataset

What does this share consist of?
The Crown Court defendant case level dataset provides data on defendant appearances in the Crown Court between January 2013 and December 2020 and is extracted from the new court management information system (Xhibit).

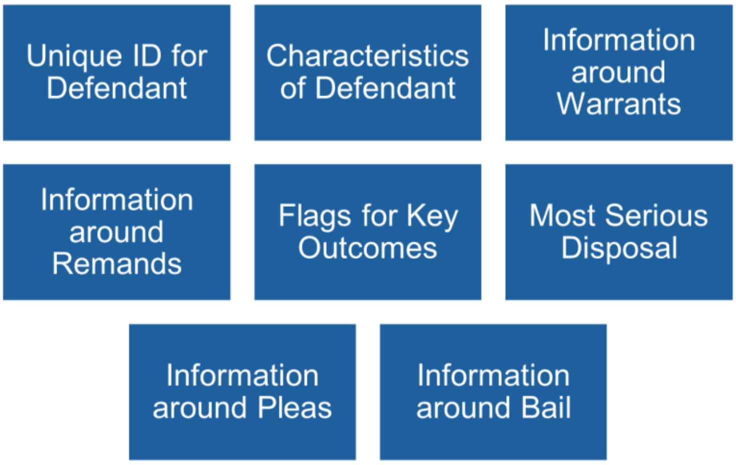


Figure 6: Descriptions of variables in the Crown Court dataset

The dataset has been deduplicated, which means multiple instances of the same person appearing before the court over the period are identified and assigned a unique identifier.

This enables researchers to establish if the same user has entered the criminal courts more than once, and the frequency, purpose and outcomes of these appearances.

The Crown Court dataset includes over 100 variables, including the dates of hearings, type of offences, arrests and the initiation date of proceedings (see Figure 7). This data explains how proceedings were dealt with (for example, whether the defendant chose trial by jury) and provides information on the final plea of the defendant. Other data include details on bail and remand.

What is the potential of this data?
Sharing this deduplicated data with accredited researchers will help develop the evidence base on the Crown Court defendants and their outcomes.

Examples of research questions that this might be able to answer:

- How do previous convictions affect sentencing outcomes in the Crown Court?
- What is the impact of earlier pleas on sentencing outcomes in the Crown Court?
- What are the associations between defendant characteristics and being sentenced to prison in the Crown Court?
- What are the key drivers of delays in the Crown Court?
- Who are the repeat users of the Crown Court? What are their characteristics? How often do they return? How long does it take for repeat users to return?

As the Data First project continues, Crown Court data will be linked to other datasets in order to address other research questions. For example, case progression from the magistrates' court to prison and probation services, and interaction with other justice services.

MoJ Data First magistrates' court and Crown Court case linking dataset

What does this share consist of?

This linking dataset will allow users to join up cases across the magistrates' and Crown Court datasets. It acts as a lookup to identify where records in the two criminal court datasets refer to the same case.

The case table gives a reference to each Crown Court record and identifies the magistrates' court case, for the same defendant, that it is judged most likely to stem from. It is generally expected that each case in the Crown Court will link back to one case in the magistrates' court. However, there could be complex instances where multiple Crown Court cases arise from a single magistrates' court case (e.g. if the individual pleads differently for different offences). Alternatively, there could be cases where there are multiple magistrates' court cases 'rolled up' into one Crown Court case.

What is the potential of this data?

Providing this linking data will allow accredited researchers to carry out new and more complex analysis, building up a fuller picture of defendants' interactions with the criminal courts, and how individual cases progress through the criminal court system.

Using this link in combination with the magistrates' and Crown Court datasets will let researchers determine that initial case X in the magistrates' court led to outcome Y once committed to the Crown Court for trial or sentencing, or on appeal.

Examples of research questions that this link could help address include:

- How do outcomes differ for triable-either-way cases heard in the magistrates' or Crown Court?
- How does an early plea in the magistrates' court impact sentence severity in the Crown Court?
- How do criminal court cases (e.g. principal offence type) change through proceedings from original receipt to the result or conviction?
- What are the differences in the characteristics of defendants appearing before the two courts for similar offences?

MoJ Data First prisoner custodial journey level dataset

What does this share consist of?

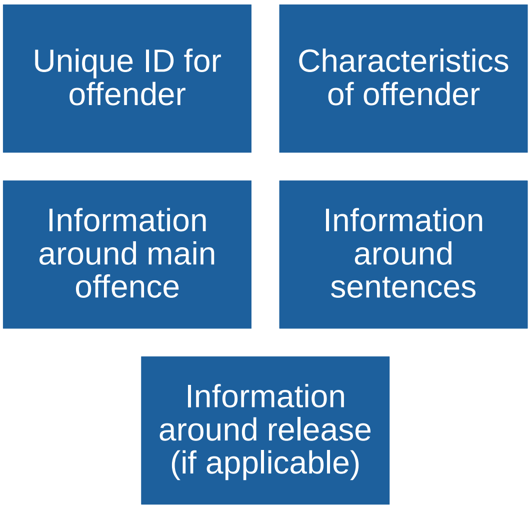

Unique ID for offender
Characteristics of offender
Information around main offence
Information around sentences
Information around release (if applicable)

Figure 7: Descriptions of variables in the prisons dataset

The prisoner custodial journey level dataset provides data on offenders’ custodial journeys in prisons and selected young offenders’ institutions in England and Wales up to 30 September 2021. It is extracted from the Prison National Offender Management Information System (P-NOMIS). Data on offenders serving custodial sentences since 2011 is expected to be complete, but sentences begun before this are included.

The dataset has been deduplicated, which means multiple records of the same person re-entering institutions over this period are identified and assigned a single unique identifier, using our data linking algorithm, Splink. This aims to improve on links already made within the prison system.

This enables researchers to follow custodial journeys recorded in P-NOMIS, providing information on movements through the system and their release (if applicable). The prisons dataset currently includes approximately 50 variables, covering information on offender characteristics, their main offence, sentence and release (see Figure 8).

This data provides opportunities for analysis in relation to prisoner demographics, factors that may impact their journey and patterns of reoffending that lead to repeat custody.

What is the potential of this data?
Sharing this deduplicated data with accredited researchers will help develop the evidence base on offenders and their outcomes.

Examples of research questions that this might be able to address include:

- What is the characteristic profile of repeat occupiers of the prison system?

- Are certain release types more likely to deter an offender from re-entering the system?
- Are offender characteristics such as ethnicity and gender associated with a variation in custodial reconviction?

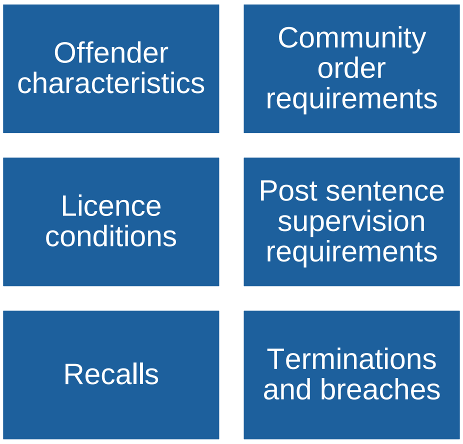
As the Data First project continues, P-NOMIS data will be linked to other datasets in order to address other research questions. For example, case progression from the criminal courts and subsequent interaction with probation or other justice services.

MoJ Data First probation dataset

What does this share consist of?
The probation dataset provides data on service users (offenders), offences, disposals, community order requirements, licence conditions and post sentence supervision requirements. It is extracted from National Delius (nDelius), a system used for the management of offenders subject to probation supervision, covering records with an event referral date from 1 January 2014, up until 31 December 2020.

There are over 100 variables spread across five separate tables within the dataset. The main table (flatfile), contains one row per an offender-event. It contains information on offender characteristics (age, gender, ethnicity, and residence), as well as offence, disposal, and sentencing information.

This flatfile table can be joined to an additional four tables. Three of these tables each represent a type of 'sentence component'. These provide information on community requirements, licence conditions, and post-sentence supervision requirements. Each of these 'sentence component' tables provide information on the length of the sentence component, as well as details about terminations.



Offender characteristics

Community order requirements

Licence conditions

Post sentence supervision requirements

Recalls

Terminations and breaches

Figure 8: Descriptions of variables in the probation dataset

The final table that can be joined to the flatfile table contains information on pre-sentence reports recorded in nDelius. This includes the type of report and the proposed requirements. Full coverage of pre-sentence reports is not available in this extract.[1]

The dataset has been deduplicated, which means multiple records of the same offender within nDelius are identified and assigned a single unique identifier, using our data linking algorithm, Splink. This enables researchers to reliably investigate offenders who are in repeat contact with the Probation Service for multiple offences.

What is the potential of this data?

Sharing this deduplicated data with accredited researchers will help develop the evidence base on offenders and their outcomes. It could be used to address priority research questions as outlined in the Ministry of Justice's Areas of Research Interest 2020, such as:

- What factors affect the likelihood of different groups receiving different sentences, including custodial, community or other court disposal sentences? How do sentencing recommendations vary by the availability of different options?
- What are the enablers and barriers to effective sentences, including community-based, alternative or short custodial sentences? Are certain types or requirements of sentences, or recommended treatment programmes, more effective for different individuals and groups?
- How has the use of non-custodial sentencing changed over time?
- What contributes to effective electronic tagging and monitoring, including GPS and radio frequency trackers, and sobriety tags, in protecting the public from harm? Are there specific groups of individuals for whom electronic tagging and monitoring is more effective?
- How can anti-social, violent, and criminal behaviour linked to alcohol and drug use be addressed beyond traditional criminal sentencing?
- What is the 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?
- How do licence period, conditions, and durations affect the potential for recalls, and what are the downstream impacts on individual outcomes and risks to public protection?
- How can short periods in custody be made more effective at reducing reoffending? What are the effects of longer custodial sentences on crime?
- How effective are rehabilitation activity requirements (RARs) and in what ways can they be improved?

[1] Where a report was not recorded in the data, it does not necessarily mean the report was **not** carried out.

MoJ Data First criminal courts, prisons and probation linking dataset

What does this share consist of?
This linking dataset will allow accredited researchers to join up persons across the magistrates' court, Crown Court, prisons, and probation datasets. The linking dataset acts as a lookup to identify where records in the various datasets refer to the same people.

The linking dataset contains rows for all records in the probation, prison, magistrates' court and Crown Court data, plus a new estimated person ID which appears alongside each instance believed to correspond to the same individual. This enables records to be grouped by individuals, and repeat appearances investigated.

What is the potential of this data?
Providing this linking dataset to accredited researchers will allow them to identify records where offenders in the prison and probation data have appeared as a defendant before the magistrates' or Crown Court. This linking dataset will enable new insights on end-to-end user journeys across the criminal justice system, including how offenders flow between key justice services.

Examples of research questions that this link could help address include:

- What factors affect the likelihood of different groups receiving different sentences?
- How has the use of non-custodial and custodial sentencing changed over time?
- Who are the 'repeat users' of the criminal justice system? What are their characteristics? How often do they return? How do outcomes change on each return to the criminal justice system?

MoJ Data First Family Court dataset

What does this share consist of?
The family court dataset provides data on people involved in family court cases in England and Wales between 1 January 2011 and 31 January 2021 and is extracted from the family court management information system (FamilyMan).

The dataset has been deduplicated, which means multiple instances of the same person appearing before the family court over the period are identified and assigned a unique identifier. This enables researchers to establish if the same user has entered the family court more than once, and the frequencies, purpose and outcomes of these appearances.

There are three tables: one that contains which people involved in family court cases are believed to be the same person (people table), the second identifies which records are for the same case (case table) and the third provides events within a case (events table).

The people table identifies each instance of the same person for each family court case they are involved in and assigns them a new unique identifier. The case table contains information on each case in the family court which are assigned a unique identifier. The events table contains one row per event within the case which can be joined to the cases table on the case. Please see the relevant data catalogue for more information.

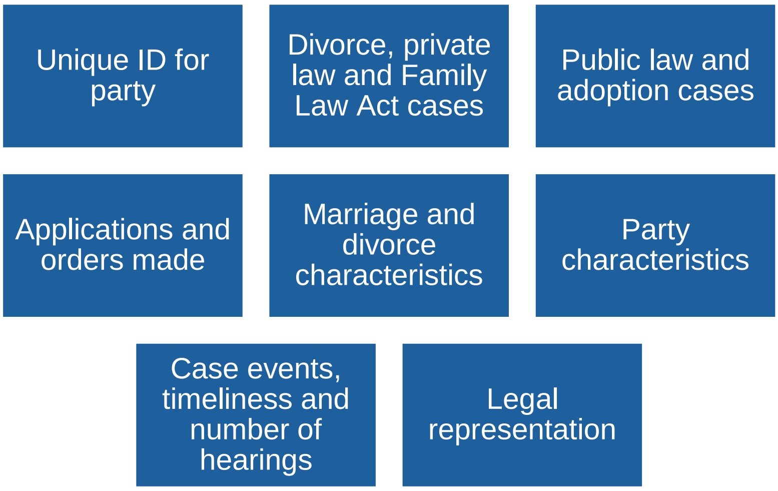

Unique ID for party
Divorce, private law and Family Law Act cases
Public law and adoption cases
Applications and orders made
Marriage and divorce characteristics
Party characteristics
Case events, timeliness and number of hearings
Legal representation

Figure 9: Descriptions of variables in the family court dataset

The family court dataset includes 76 variables, including dates around proceedings, information around applications and legal representation the parties involved with the case (see figure 10).

What is the potential of this data?
Sharing this deduplicated data with accredited researchers will help develop the evidence base on family court users and their outcomes.

Example of research questions that this might be able to answer include:

- What is the nature and extent of repeat use of the family court?
- Who are our repeat users? What are their demographic characteristics? Do the same parties return to the court in different contexts?
- How do public and private law cases overlap? What are the characteristics of these users? What are their pathways through the court and the outcomes for these groups?

MoJ and DfE data-share

What does this share consist of?
The MoJ-DfE share provides data on childhood characteristics, educational outcomes and (re)-offending. The shared information consists of data on the educational characteristics of young people (from DfE), linked to data on their interactions with the criminal justice system (from MoJ).

The data relates to those offenders with at least one record from 2000 or later, who were on the Police National Computer (PNC) at the end of 2017 and were matched to individuals on the National Pupil Database (NPD). Only offenders who were born on, or after 31 August 1985 were matched, because earlier groups do not have a realistic chance of matching. The earliest year shared will cover those aged 16 during the 2001/02 academic year, the oldest group likely to be present in the NPD.

This data share includes 20 DfE datasets, including data on academic achievement, pupil absence and pupil exclusions. It also includes 11 MoJ datasets, including data on offenders' criminal histories, court appearances and time in prison. Each dataset has a unique ID variable that can be used to link across the datasets.

Copies of this share are held by both DfE and MoJ, enabling it to be contextualised with data from the wider justice and education systems. The shared information, together with wider education data, is available to researchers through the ONS SRS. If researchers require wider justice data in addition to the share, they can apply for access through the Justice MicroData Lab. Although the share itself was not developed within Data First, the programme includes provision to support users to make best use of it.

What is the potential of this data?
This data-share has been undertaken for the purpose of increasing understanding of the links between childhood characteristics, education outcomes and (re)-offending. Sharing this data with accredited researchers will help develop the evidence base on understanding the relationships between educational and criminal justice outcomes and the drivers of offending. It will assist in identifying the population that requires support through early intervention and evaluating these projects to understand whether they are effective.

Examples of research questions this may be able to answer:

- Is there an association between particular interactions with the education system and offending, and if so, is one of these factors typically the driver?
- How are the relationships between educational and criminal justice outcomes impacted by demographic factors?
- Are interventions to prevent Serious Violence effective (through use to generate control groups)?

Data Catalogues

What is a data catalogue?
A data catalogue is a collection of descriptive data (metadata) that aids researchers in understanding what data is available and helps them to find the data that they need. Researchers can use a data catalogue to understand what potential research questions a dataset can answer, and identify the variable names required in their application form to seek access to the data.

The data catalogues for Data First include metadata around each variable that is being shared, including their quality, formatting, type and description (see Figure 10), which will aid researchers in exploring and analysing the data.

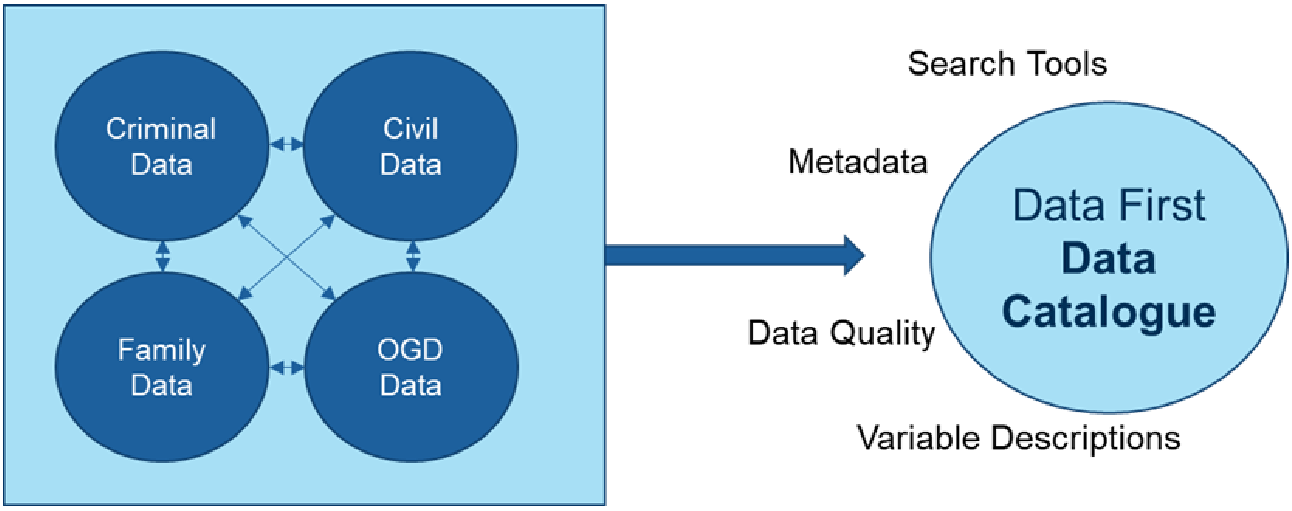


Figure 10: Demonstration of the link between the datasets and the Data Catalogue

Possible values of categorical or scale variables are also given, together with lookups to interpret any values stored as code sets. Changes in methodology or definition will also be included and highlighted where possible.

The catalogues also seek to provide researchers with more detailed information on the scope and coverage of records in the shared datasets as well as their processing and data sources.

Which Data Catalogues are available?

- MoJ Data First magistrates' court defendant case level dataset
- MoJ Data First Crown Court defendant case level dataset
- MoJ Data First prisoner custodial journey level dataset
- MoJ Data First Family Court dataset
- MoJ Data First probation dataset
- MoJ Data First criminal courts, prisons and probation linking dataset

Information will be added to the published data catalogues as more datasets are released through Data First.

What are the uses and limitations of the data?

Data quality

Administrative data

Administrative data refers to information which was originally collected for non-statistical reasons, to enable the delivery of a public programme or service, or to maintain records. 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.

The administrative data that is being used by Data First was originally collected for the purposes of administering the justice system (or other government services), such as to enable the processing of court cases and the running of prison and probation services.

Data sources at MoJ

The justice sector has a complex landscape of management information systems, designed independently to meet the needs of its diverse organisations and functions.

For example, while courts and tribunals have been managed by Her Majesty's Courts and Tribunals Service (HMCTS) since 2011, many operational systems pre-date this. The data processing requirements for handling a child supervision order, a civil money claim or a criminal trial differ, with different information in each case being passed into centralised systems. Many court processes are handled locally by individual courts, and frequently involve paper forms and documents, although the ongoing Court Reform Programme aims to streamline and modernise data collection and processing. Some data collected at source are therefore never centrally compiled once their immediate use is met.

The criminal justice system further involves data collection by a network of departments and agencies, from police forces and Crown Prosecution Service (CPS) to Her Majesty's Prison and Probation Service (HMPPS). Creation of a Common Platform digital infrastructure which will deliver a single online system and remove the need for manual handling of documents, duplication of process and the re-keying of information is a major government IT project, the need for which is clear because of the lack of interoperability between current systems. Bringing together and structuring data on offender journeys to recognise a coherent user journey from start to end will require substantial groundwork.

Other information management systems that provide information on offender journeys through the criminal justice system that may be used as data sources in our upcoming data shares under Data First include:

- The Police National Computer (PNC) – a large administrative database containing information about police cautions and court convictions held on individual offenders in England and Wales. The PNC is regularly updated as new information about particular individuals becomes available.
- Prison population data is drawn from the prison National Offender Management Information System (p-NOMIS). Whilst the PNC provides details on offences committed and sentences given to offenders, snapshot data taken from p-NOMIS provide information on the number of offenders currently serving custodial sentences. Since 30 June 2015, due to improvements in the systems used for processing data extracted from p-NOMIS, more detailed information about the prison population has been available.
- The national Delius (nDelius) system records the flow of offenders released from prison and starting community sentences. This is for all adult offenders discharged from custody (determinate and indeterminate sentences) and for those managed in the community.
- The Offender Assessment System (OASys) was introduced in 2001 and built on the existing ‘What Works’ evidence base. It combines the best of actuarial methods of prediction with structured professional judgement to provide standardised assessments of offenders’ risks and needs, helping to link these risks and needs to individualised sentence and risk management plans.

Impacts of data quality on research

While MoJ holds a wealth of administrative data, there is little harmonisation of the fields collected across different areas, creating inconsistency in data definitions, data formats and values. Data collection covers what is necessary for operational purposes, meaning that data may not be of the desired quality or comprehensiveness, or be consistently available, to address important academic questions.

Over time, changes to management information systems, processes and policies may have introduced breaks in time series that could affect analysis and interpretation. 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. 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 First must identify appropriate versions of data from sometimes complex existing pipelines to release for academic purposes and determine the most suitable variables to include. While certain data have long been processed internally, for example for release in

Official Statistics publications, the programme will include fields that have not been subject to extensive internal analysis and for which less is known about limitations.

Although MoJ has taken the necessary steps to ensure that the data within these shares is suitable for research purposes, and to provide data catalogues to aid with understanding of definitions and methodology, the researcher must be mindful of its source. Merging and processing datasets from different systems and organisations is a complex task and methods of best practice are still being developed and established. In releasing this resource to the wider research community, we hope to increase our understanding of the source data. By collaborating and sharing experience and expertise we can improve our assessment of its strengths and weaknesses in addressing research questions as the programme develops.

Data linkage

What data is being linked?

The internal administrative datasets MoJ are bringing together as part of Data First each represent an interaction of a 'user' of the justice system (e.g. 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's short-term aim is to provide a unique ID for researchers working with one dataset and then to produce a synthetic unique identifier that can be used to link records reliably across justice system datasets.

We are linking records only at the level of an individual, to allow analysis of a person's journeys through the justice system. We 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).

Data linking process

Without a unique personal identifier, we rely on comparing other identifying information, 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[2]) 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-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.

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

Assigning a match

For the magistrates’ court defendant case level dataset, Data First has deduplicated 15 million records using the Splink package, calculating match scores for 160 million record comparisons. Records 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 can then be added to any datasets used by researchers.

Without a labelled training dataset (where we know 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’ court data, we have clerically scored tens of thousands of record pairs to simulate this ideal labelled data. Providing detail on matching probabilities is not straightforward as these are calculated pair-wise 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 match, 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. We make no attempt to define a canonical record of personal information (determine which is the correct/current value of name, address etc.) but just 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 (e.g. 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 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 more unusual 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.

How is the data protected?

Data privacy and security

De-identifying the data

Personal identifiers are used for linking and the deduplication of individuals in the data but will **NOT** be shared as part of Data First. This includes variables such as:

- Forenames
- Surname
- Date of birth
- Home address

These are used to generate a meaningless unique person ID.

Replacement values are generated for internal system IDs to prevent direct linkage back to the raw data source. These include:

- Defendant ID
- Case ID
- Row ID

Some special category related 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, for example:

- Sex/gender
- Ethnicity
- Age in years

Is the SRS Safe?

The ONS SRS uses 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.
- 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.

Legislation

The Data First project 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).

Any researcher who wants to access the data through the SRS must be accredited under the DEA and the research project must have been approved by the data supplier(s) and the Research Accreditation Panel. 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.

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.

Applying for data access

Accessing Data First data

Access is available to internal linked data, and external linked data where agreed, through the ONS Secure Research Service (SRS) and in certain cases, the Justice MicroData Lab (JMDL). From the researcher's perspective there are two steps needed to access the data:

1) Become an accredited researcher with ONS (and, for JMDL, complete Safe Researcher or SURE training).
2) Apply for access to the specific data required for a research project from relevant data owners using the MoJ and HMCTS combined form.

Becoming an accredited researcher

Anyone who needs to access the SRS will need to be accredited under the ONS accredited researcher scheme. We recommend that you begin this process as soon as possible and in parallel to any research application. 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 11).

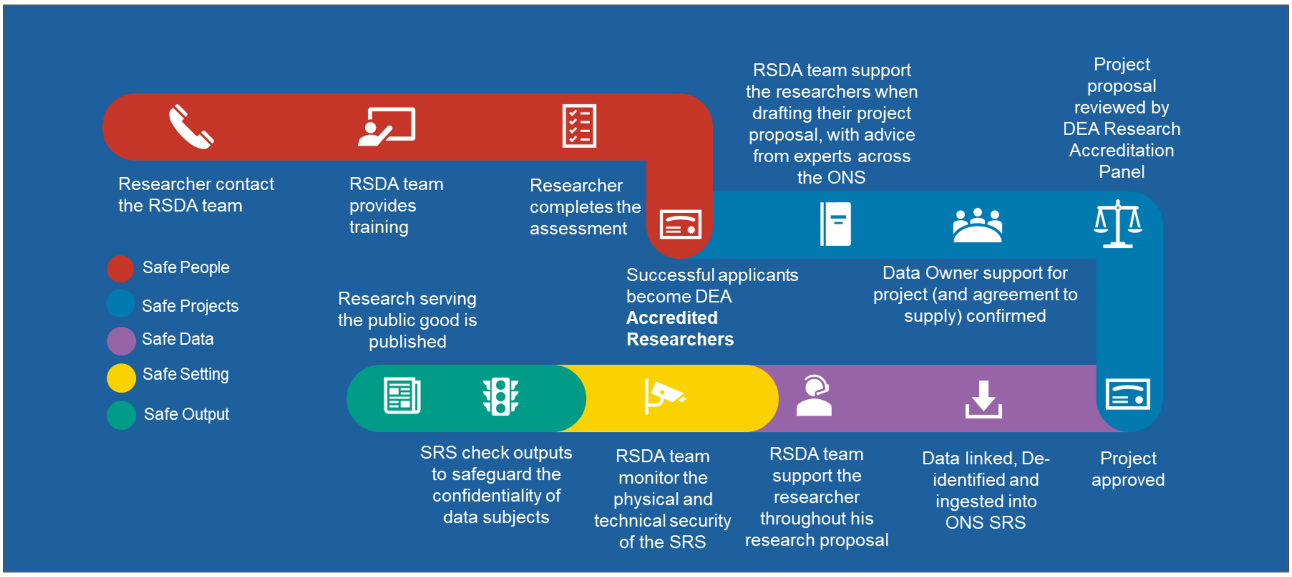

Researcher contact the RSDA team
RSDA team provides training
Researcher completes the assessment
Successful applicants become DEA **Accredited Researchers**
RSDA team support the researchers when drafting their project proposal, with advice from experts across the ONS
Data Owner support for project (and agreement to supply) confirmed
Project proposal reviewed by DEA Research Accreditation Panel
Project approved
Data linked, De-identified and ingested into ONS SRS
RSDA team support the researcher throughout his research proposal
RSDA team monitor the physical and technical security of the SRS
SRS check outputs to safeguard the confidentiality of data subjects
Research serving the public good is published
● Safe People
● Safe Projects
● Safe Data
● Safe Setting
● Safe Output

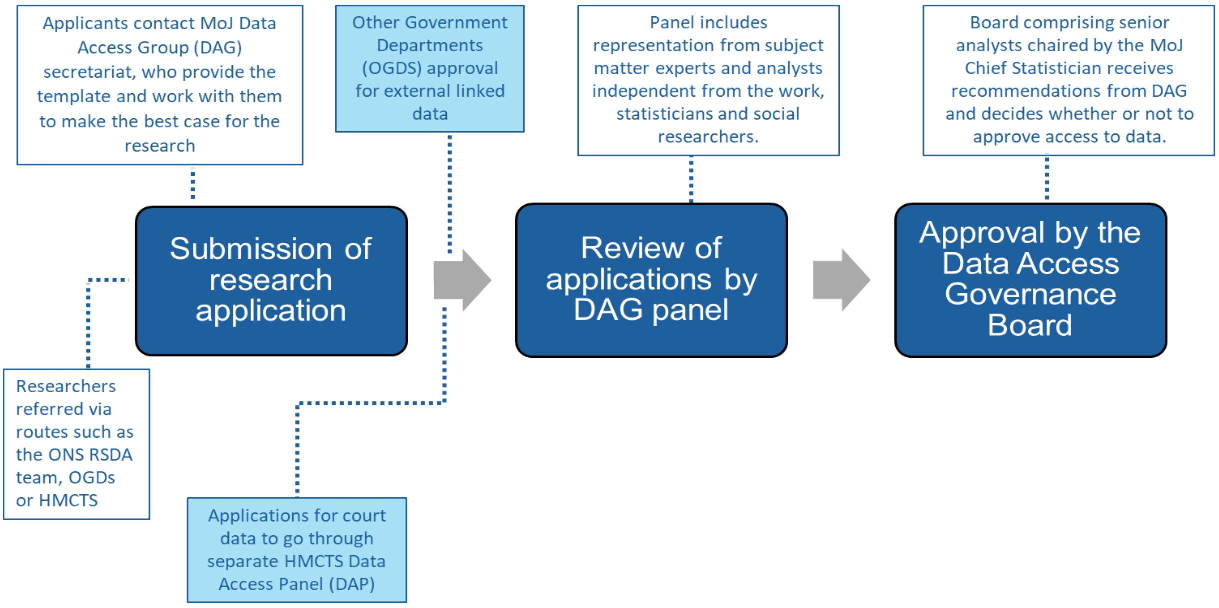
Figure 11: A walkthrough of a researcher's accreditation and project development journey

The RSDA team creates and looks after all the training requirements for those looking to become accredited researchers, from booking venues to sending out invites and ensuring the Statistical Support Team attends to deliver the training. Once training is complete the RSDA team advises and supports the researchers who have passed the course and achieved their accredited status on the drafting of research proposals, providing them with advice from ONS experts.

Data First research governance processes

As stated in Figure 11, approval from the relevant data owner must be gained to access data for a research project. To gain access to MoJ-specific data-shares made available through Data First, accredited researchers must apply to MoJ's Data Access Governance Board (DAGB) for approval, with proposals first assessed by the Data Access Group (DAG) on DAGB's behalf.

The governance process for seeking approval from the DAGB is outlined in Figure 12.



Applicants contact MoJ Data Access Group (DAG) secretariat, who provide the template and work with them to make the best case for the research

Other Government Departments (OGDS) approval for external linked data

Panel includes representation from subject matter experts and analysts independent from the work, statisticians and social researchers.

Board comprising senior analysts chaired by the MoJ Chief Statistician receives recommendations from DAG and decides whether or not to approve access to data.

Submission of research application

Review of applications by DAG panel

Approval by the Data Access Governance Board

Researchers referred via routes such as the ONS RSDA team, OGDs or HMCTS

Applications for court data to go through separate HMCTS Data Access Panel (DAP)

Figure 12: The process for gaining the approval of the DAGB for access to a Data First share

If applicants are seeking access to court data, approval from the HMCTS Data Access Panel (DAP) will also be required. Approval will be required from our partner departments for external data-shares.

Accredited researchers submit a research application stating their research proposal, how the potential findings could be beneficial, their methodology and information around ethical considerations and data security. They must include information about the data items they wish to access and analyse in the project, with justification.

A single application form combining requests to both the DAGB and the DAP is available on gov.uk. This form combines requests for access under Data First and other MoJ and HMCTS data. A separate guidance document directs researchers to which sections to complete.

This application will then be reviewed by analysts and data experts within MoJ and its partners who will judge whether the research proposal meets the required criteria. The panel explicitly consider whether the research proposal is ethical (including its potential impact on data subjects); whether the data is necessary to address the research questions; that data protection concerns are addressed; and the overall benefit of the research. Their recommendations will be scrutinised by a board chaired by MoJ's Chief Statistician who decide if the researcher will be granted access to the data they requested and can carry on with their research journey.

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. We will also consult with a representative user panel, bringing together organisations that represent justice system users, to ensure public acceptability of the work.

A brief summary of successful applications to access the data will be published on gov.uk.

ONS Secure Research Service

Access to Data First datasets is currently only available through the ONS SRS, which is an Accredited Processor under the DEA (2017). The SRS has substantial data expertise, especially in data management, metadata, and the checking of outputs before they leave the centre.

Once approval has been gained from the data owner to access data for a specific research project, as outlined above, they liaise with the RSDA team, and the project goes to the UKSA. If approved, the data extract required can be prepared so the project can start. The RSDA team will continue to provide guidance and checks for the researcher until publication.

How do I use the ONS SRS?

How to access the Server

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.

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. Below is a list of the software that is currently available to use for researchers:

Software	Version
STATA	14
SPSS	24
SAS	9.3
R for Windows	3.5.2
ML- WIN	3.02
QGis	2.18.19
Microsoft Office Professional Plus	2013 15.0
7zip	18.01
Anaconda3	5.1.0
Adobe Acrobat Reader DC	18.011
Notepad++	7.56

Software	Version
Winzip	18.5
ArcGIS	10.4.1
R Studio	0.99.903
Jupyter Notebook	5.1.0
Qtconsole	5.1.0
Spyder	5.1.0

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.

Further Information

General contact and enquiries	datafirst@justice.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