

ONS policy for safeguarding data whilst managing Admin Data Research Network projects

May 2015

Background

The Administrative Data Research Network (ADRN)¹ is an initiative funded by the Economic and Social Research Council with the aim of making linked, de-identified data from administrative sources available for research in a secure environment.

ADRN services are delivered by four research centres, one in each country of the UK, led by the University of Southampton, the University of Edinburgh, Swansea University and Queen's University in Belfast. The work of the four research centres is co-ordinated by the Administrative Data Service (ADS).

The network is governed by the ADRN Board located in the UK Statistics Authority. To access the services provided by any centre, research projects are assessed by an independent approvals panel, and the researchers are also accredited.

The Administrative Data Research Centre for England (ADRC-E) is run collaboratively by the University of Southampton, University College London, London School of Hygiene and Tropical Medicine, Institute for Fiscal Studies and the Office for National Statistics (ONS). We contribute to the ADRC-E through our expertise in securely managing identifiable and sensitive data, linking together administrative data, and providing a secure environment where researchers can work with the de-identified data needed for their research.

About this paper

This paper explains how ONS will safeguard the confidentiality of personal information throughout its involvement in ADRN projects.

¹ For more information on the ADRN visit www.adrn.ac.uk

Table of Contents

1	Summary.....	2
2	Introduction	4
3	Research Project Approval.....	5
4	Transferring Data Securely.....	6
4.1	Security Measures	6
5	Creating Linked Datasets Securely.....	7
5.1	Security Measures	7
5.2	Anonymising the identifier files	7
5.3	Linking the identifier files	8
5.4	Matching the attributes files.....	8
5.5	Pre-release assessment of the linked attribute file	9
6	Transferring Datasets to a Secure Environment.....	9
6.1	Security Measures	9
7	ADRC-E researcher access in the ONS secure environment.....	10
7.1	Procedural security measures	10
7.2	Physical Security Measures in the VML server room.....	11
7.3	Digital Security Measures.....	11
8	Conclusion	11
Appendix A: Operating procedures of the linkage facility		12
Appendix B: Detailed specification for the cryptographic functions in use		24
Appendix C: Glossary of Terms used		28

1 Summary

The Administrative Data Research Centre for England (ADRC-E), and the other national Administrative Data Research Centres within the UK, are tasked with meeting the objectives of the Administrative Data Research Network (ADRN) by facilitating research projects approved by the network.

ADRN projects require the legal and secure acquisition of individual-level administrative datasets which are securely linked creating a de-identified research dataset available to approved researchers in a secure environment. ONS may be involved in one or more of these steps but primarily we will provide secure data linkage, and a secure environment where approved researchers can access only the linked data relevant to their approved project.²

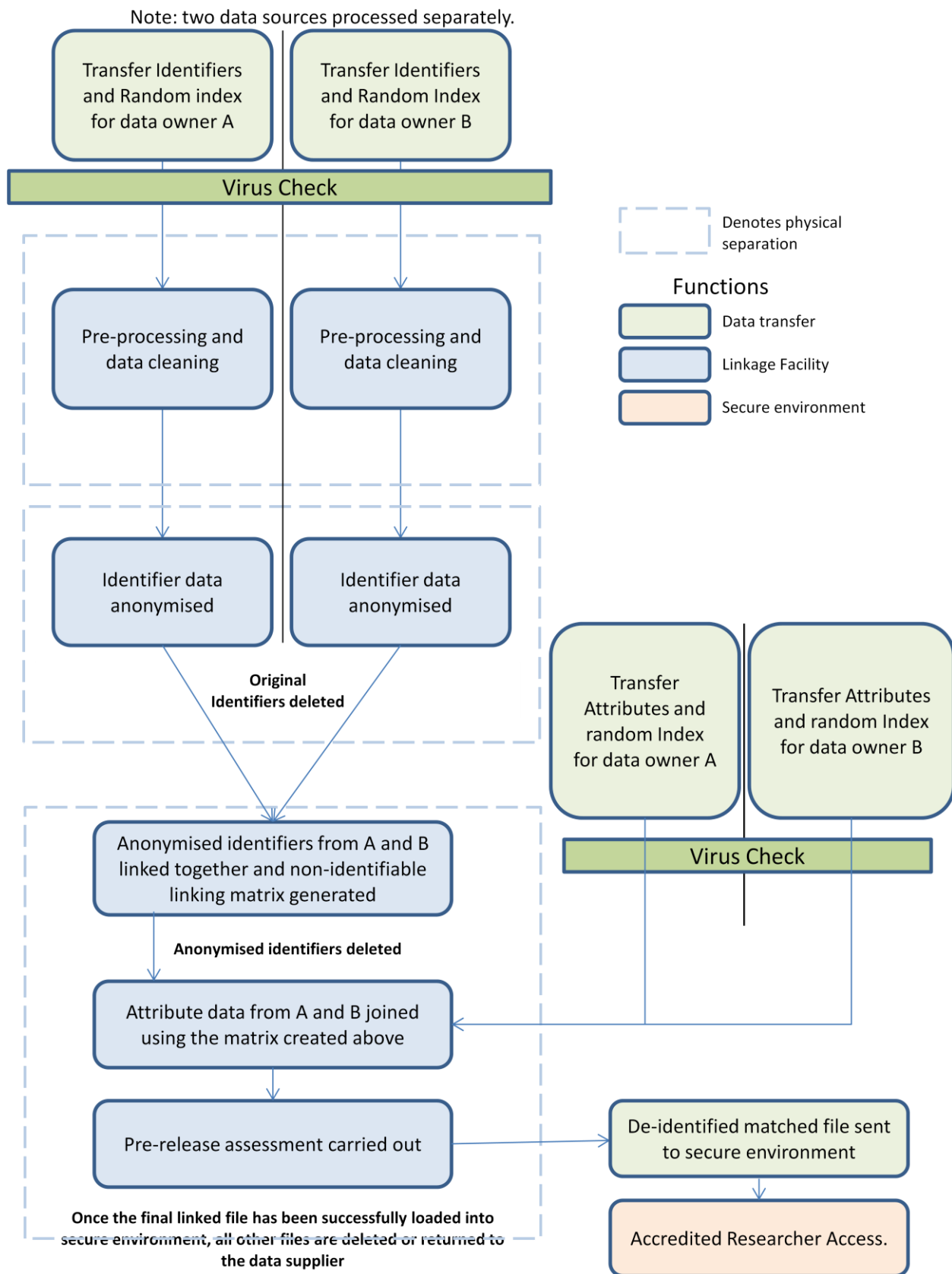
ONS has a legal obligation to protect the confidentiality of all the information it holds, including datasets we handle during ADRN projects. We have designed systems, procedures and methods for our contribution to the ADRC-E that minimise the risk of inadvertent or deliberate disclosure of any personal information. In particular, full account has been taken of the risks associated with:

- transfer of administrative data from data owner to ONS
- securing access to the data to be linked
- the management of both the identifying data and attributes contained within the administrative data sources
- the linkage process
- the identification of population subgroups within the linked data
- transfer of the de-identified linked file to a secure environment
- secure access by researchers at the secure environment.

In addition, ADRN policy requires strict separation between those linking the administrative data and those providing the secure environment where researchers access the linked data for their project. Often, ADRCs achieve this by arranging for a different trusted organisation to securely link the data, while providing the secure environment for the researchers directly. However, because ONS may provide both functions for an ADRN project, we employ strict procedures and separation of personnel when completing work in each area to ensure this separation is achieved.

² Our ADRN involvement builds on existing experience, infrastructure and capability in safeguarding data, in particular the [Beyond 2011 programme](#), and the [Virtual Microdata Laboratory](#)

Figure 1: Data processing steps covered in this document



2 Introduction

The Office for National Statistics is working with the University of Southampton, University College London, London School of Hygiene and Tropical Medicine, and the Institute for Fiscal Studies to run the Administrative Data Research Centre for England (ADRC-E).

The ADRC-E and other national Administrative Data Research Centres within the UK are tasked with meeting the objectives of the Administrative Data Research Network (ADRN) by facilitating research projects the network approves. To do this, individual-level administrative datasets must be legally and securely acquired, securely linked, de-identified, and then placed in a secure environment where only the relevant approved researchers can access the data.

ONS' role within the ADRC-E is to provide a secure environment where approved researchers can access the linked data for their project.³ In addition, we provide the trusted linkage function on behalf of the ADRC-E. To protect confidentiality while undertaking these tasks, we will:

- ensure secure transfer of administrative data approved for use by ADRN projects from the organisation that owns the data to ONS⁴
- maintain strict separation between the identifiers and attributes for all ADRN project datasets
- provide a linkage facility and service where the identifiers from each dataset are anonymised and linked together
- delete the identifiers once a linkage matrix has been created, and use this to link the attributes from each dataset
- assess the linked attribute dataset for disclosure risk, and where necessary adjust the dataset to protect against disclosure
- disseminate the linked data via a secure environment, our Virtual Microdata Laboratory (VML), where the relevant approved researchers can access only the linked data for their approved project

These data processing steps are shown in Figure 1 (above). ONS also provides a co-ordination function to ensure our involvement is consistent with the principles that drive the ADRN. We will regularly review our operational procedures as wider ADRN policies are developed, updating them as required.

This rest of this document explains the lifecycle of an ADRN project with a focus on how ONS safeguards the data when we are responsible for a stage in this process on behalf of the ADRC-E. Other parts of the ADRN will release their own statements about safeguarding data.

More specific technical information about our linkage service (appendix A) is provided.

³ Not all ADRC-E projects will use the ONS linkage service. Similarly, other secure environments for ADRC-E research are available at the University of Southampton, and the Farr Institute in UCL.

⁴ ONS will also be a data owner providing data for some ADRN projects. The policies outlined in this paper apply equally to ONS and non-ONS data.

3 Research Project Approval

Administrative Data Research Network (ADRN) projects are initiated by a lead researcher submitting a project proposal to the Administrative Data Service (ADS), along with an approved researcher application form for each researcher in the team.

The ADS and/or designated Administrative Data Research Centre (ADRC) will assist with the preparation of the form, and may conduct initial investigations with methodologists, legal and data specialists to prepare a view on the feasibility of the proposed project. The ADS will submit a detailed proposal to an Approvals Panel (AP) that is independent of the operations of the ADS and ADRCs and includes lay membership.

The AP will scrutinise the proposal around the following five aspects:

- Is the project feasible?
- Have any relevant privacy implications been addressed?
- Has the project been through a formal ethical review?
- Is there potential public benefit?
- Is there demonstrable scientific merit?

The AP will make one of three decisions:

- 1) Approve
- 2) More information needed
- 3) Reject

If an application is rejected, a description of why it failed is provided.

If a project is immediately or subsequently approved, the ADS will commence formal negotiations with the data owners from whom the project is requesting data. This will be an iterative process and may involve staff from the relevant ADRC where they have existing relationships with the data suppliers.

Data owners are ultimately responsible for the security of the data and must determine whether supply of the data is lawful. They will also decide whether preparing and extracting the data carries any cost they would have to pass on to the researcher in order to meet the request. The data owner may impose conditions on where the linking can occur and which secure environment must be used based on their risk appetite and the accreditation status of the linkage facility or secure environment.

The ADRN has given an undertaking that identifiers, that is information such as name, address and date of birth, will be kept separate from the attributes that form the basis of the research, such as educational attainment, benefit claimant periods, or car ownership. The remainder of this paper will refer to the two files as the identifiers file and the attribute file.

4 Transferring Data Securely

When ONS provide the data linkage service for a project assigned to the ADRC-E, the data must first be transferred to ONS.

All data handling carried out by ONS complies with Cabinet Office Information Assurance policy. Regardless of the original source of a dataset, even where it is being transferred within a single ONS building, data are transferred into the ONS linkage facility on CESG-approved⁵ encrypted media, with encryption passwords and/or tokens controlled by ONS Security Managers⁶. Under no circumstances are personal data transported without appropriate levels of encryption in place. Some data providers will choose to transmit their data using the Public Sector Network (PSN), rather than on encrypted media. Appropriate encryption for the data will be employed on all files transmitted in this manner and on receipt will be transferred onto CESG-approved encrypted media.

The sensitivity of the information in each dataset is identified by the data owner and the level of sensitivity is used to determine the level of protection that is assigned to the data when it is transferred. For more information, please see the [Cabinet Office Security Policy Framework](#).

4.1 Security Measures

Data transfer is one of the concerns reported in a recent [public engagement on the ADRN](#). Under the Data Protection Act 1998 ONS will act as a data processor for individual organisations (for example, government departments) who are willing to provide data for research purposes, but the responsibility for the data and its safe transfer to ONS remains with the data owner. However, where appropriate, ONS will facilitate by providing the CESG-approved encrypted media for the transfer and more generally, there are minimum standards of security that ONS is willing to accept in any transfer:

- **Physical** – Secure Disk Drives are used for physical transfer. These are password protected devices and approved by CESG for Government data transfer. The Public Sector Network (PSN) is a secure network for the transfer of files, accredited by the CESG Pan Government Accreditation service. Storage of devices containing data will be kept to a minimum, and will only be in a secure safe.
- **Procedural** – All files in transit will be encrypted in accordance with CESG policy, and the policies of the owning department. Passwords will only be shared with nominated personnel (Security Managers). Files will be separated so that attribute data and identifier data are never transferred together.
- **Personnel** – All operations carried out during data transfer and load are done by ONS Security Managers holding Developed Vetting⁷ clearance.
- **Technical** – the computer that receives encrypted files from the PSN is physically disconnected from any ONS network – it only accesses the PSN.

⁵ That is, media approved by the UK Government's National Technical Authority for Information Assurance (CESG - Communications Electronics Security Group)

⁶ ONS Security Managers are in a separate management team, independent from the ONS ADRC-E personnel

⁷ The most detailed and comprehensive form of UK vetting, required for sensitive jobs.

5 Creating Linked Datasets Securely

Once the data have been securely transferred to ONS, they must be securely linked together. ADRN policy requires that for any given dataset, identifying information such as name and address are kept separate from attribute information. The identifier files are used to determine how two datasets link together at record level. This allows creation of a non-identifiable matrix that can be used to link the two attributes files together, without the need for the identifiable information from the identifier files.

In addition, ONS has implemented other procedures that ensure that high levels of anonymity and privacy are maintained.

5.1 Security Measures

The ONS linkage facility has been designed specifically to address any privacy and security concerns that may arise when datasets are linked. Robust security controls are in place to ensure the safety of all information and to ensure that confidentiality is protected.

In summary, the security controls in place include:

- **Physical** – all linkage takes place behind high-security doors in a secure physical environment.
- **Procedural** – all data acquisition, import and export processes are subject to strict procedural controls.
- **Personnel** – only authorised employees holding Security Check clearance are permitted to enter the environment, and all access is recorded, monitored and audited by ONS Security Managers on a regular basis, through regular review of technical, procedural and CCTV records.
- **Technical** – the linkage facility is fully isolated from all other systems and networks. Within the environment, technical safeguards exist to ensure only authorised work can take place, and “unusual” activity is detected, assessed and acted upon. Electronic devices, (including mobile phones), software or connections are not permitted in the environment under any circumstances, and protective measures are in place to enforce this policy.

Finally, all processes for the storage and retention of data take account of obligations in the Data Protection Act 1998, the Statistics and Registration Service Act 2007, Government data security and handling standards and the requirements of individual data suppliers. Once the data have been imported to the linkage facility, the physical media on which the data arrived are either stored securely and separately in a locked safe accessible only to Security Managers, or returned to the data supplier. Once ONS has completed the linkage work all data are returned to the supplying organisation or destroyed using appropriate techniques.

5.2 Anonymising the identifier files

Identifier files include a randomly generated unique index for each record. The same number is assigned to the attributes for that record in the separate attributes file. Before identifier files are linked together, ONS employs a process whereby the identifiable data⁸ in each identifier are pseudonymised. In line with the Information Commissioner’s recent [Code of Practice for](#)

⁸ That is, data by which an individual could be identified, such as name, address and date of birth, and any other unique identifier which could be used to uncover such information.

[Anonymisation](#) this means that all uniquely-identifiable fields within each dataset are cryptographically hashed⁹. The random index is not hashed.

Hashing is undertaken by one team on one IT system, while the linkage is completed by another team on a completely separate IT system. The hashing system is physically disconnected from all other infrastructure while the hashing process takes place.

As a result, the linkage team cannot see any unhashed personally identifiable data, only the random index. Only fully automated linkage methods are used to link one identifier file with another at the record level. On the other hand, the hashing team do not work with more than one identifier file at a time. Once an identifier file is hashed and made available for linking, all remaining data are securely and permanently deleted before the next identifier file is hashed.

Appendix B includes an illustrative example of what data looks like, and therefore what the linkage team see. We do not employ clerical matching as this is only possible when unhashed identifiers are available.

The fields that are hashed for all identifier datasets are¹⁰:

- First Name(s) and initial(s)
- Surname(s)
- Date of Birth
- Postcodes

5.3 Linking the identifier files

Once all identifier files for the datasets to be linked have been hashed, they are linked using our linkage methodology. This has been robustly tested, the accuracy is well understood, and it is fully explained in appendix A. The linking process ties together the random indices from each dataset to produce a matrix describing which attribute data from one dataset relates to which attribute data from another. Once this matrix has been produced, the hashed identifiers are not required to complete the final stage of the matching, and they are deleted.

5.4 Matching the attributes files

The attribute data are loaded into the linking facility once the hashed identifier data have been deleted, so the attributes and identifiers never co-exist in any system. The attribute files for each dataset are matched together using the non-identifiable indices matrix produced at 5.3. This process is automatic.

There will be some records from each dataset that have not been matched. The attributes data for the non-matched records will be risk assessed and supplied to researchers when possible. However these records may present a disclosure risk and in such cases will not be supplied, depending on the requirements of the data owner. As a minimum, researchers will be provided with non-disclosive aggregate data to help account for any bias in the linking process.

⁹ It is important to recognise a distinction in the use of this terminology, and therefore the processes in place. We refer to *encryption* as a two-way process which can be reversed – an encryption key is stored and can later be used to ‘undo’ the encrypted values, restoring the original data. *Hashing* is a one-way process that is irreversible – once the hashing algorithm is applied it is not possible to get back the original information in the linkage facility.

¹⁰ This is the default set of identifiers for personal data, other configurations are possible.

5.5 Pre-release assessment of the linked attribute file

In order to check for any privacy or data security issues that might have occurred during the linking and matching process, a full risk assessment is carried out on the linked attributes file.

Any issues discovered during this assessment will be resolved in consultation with ADRC-E co-ordinators, data owners and researchers as appropriate. This may entail reducing the granularity of the data provided, for example collapsing age into 5-year age bands.

The risk assessment will result in an agreed Business Impact Level (BIL) and Protective Marking (PM) for the linked de-identified dataset which will be used by the secure environment as authority that the matched dataset can be imported. The Director for the Admin Data Division in ONS will be the Information Asset Owner for the linked dataset within ONS. They are responsible for its security while it resides within ONS, and for assignment of BIL and PM. However, the respective owners of the original datasets that have been linked still own their respective data, are joint owners of the linked file, and have ultimate control of those data.

6 Transferring Datasets to a Secure Environment

The linked attribute dataset will be protected at all times whilst in transit from the linkage facility to the relevant secure environment. This may be the ONS secure environment, but could also be the other ADRC-E secure environments at the University of Southampton or Farr Institute at UCL.

Ultimate responsibility for safe transit is still held by the data owners, so ONS will adhere to any requirements specified by the data owners in making this transfer, and we will provide the CESSG-approved encrypted media for the data transfer.

Where the linked dataset contains ONS data, all activities in transfer are governed by the Statistics and Registration Service Act, 2007. Authority to transfer will be made by the ONS Microdata Release Panel (MRP) in accordance with instructions from the data owners whose data make up the linked dataset.

Where the linked dataset is made up of data entirely from other government departments (or their agencies), authority for transfer will be made by the Director for the ONS Admin Data Division, in accordance with instructions from the data owners whose data make up the linked dataset.

6.1 Security Measures

- **Physical** – Secure disk drives used for physical transfer. These are password protected devices and approved by CESSG for Government data transfer.
- **Procedural** – All files in transit will be encrypted in accordance with CESSG policy, and the policies of the owning department. Passwords will only be shared with nominated personnel (Linkage or Security Managers). Authority to transfer will be either from the MRP or the Director for the Administrative Data Division.
- **Personnel** – Extraction of data files from the ONS ADRC-E Linkage Facility can only be done by Security Managers holding Vetting clearance. Receipt and loading of the data files will be in accordance with the ADRN approved Safeguarding policies of the secure environment.
- **Technical** – Files will only be transferred to and from the secure environment infrastructure, accredited to the appropriate level, using secure disk drives.

7 ADRC-E researcher access in the ONS secure environment

There are several secure environments where researchers may access the linked the data for ADRC-E projects:

- The ONS Virtual Microdata Laboratory (VML)
- A secure environment at the University of Southampton
- A secure environment at the Farr Institute, University College London.

The secure environment used will depend on factors including researcher preference and any restrictions specified by the owners of the linked datasets.

When the ONS VML is used, the data are:

- stored in a data centre based within the UK mainland that has been accredited as secure by HM Government
- only transferred into the VML from other ONS systems in an encrypted format
- always be de-identified, with no directly identifiable variables (e.g. name, address, date of birth)
- only accessed by researchers who have the necessary permissions to analyse them
- only accessed from a secure, approved, location, never from the public internet.

In addition ONS VML staff:

- have appropriate security clearance
- only allow access by researchers accredited by the ADRN, and only to the data for their approved project
- check all outputs from the VML, to ensure they are not disclosive to individuals, households or businesses, before making available to researchers for dissemination or publication.

The ONS VML is based on long standing infrastructure and the VML team has expertise and a proven track record in safely disseminating data for research while protecting confidentiality. The VML is, and will continue to be, used for non-ADRN research. Where any detailed procedures that apply only to ADRN projects are developed to meet evolving ADRN policy, they will specified in an appendix to this paper. In addition, ONS VML policy will be reviewed annually by the VML Security Manager, and signed off by the ONS Senior Information Risk Owner, to ensure that it remains up-to-date, comprehensive, and to identify any additional measures required.

7.1 Procedural security measures

Lawful access for ADRN researchers in the VML will be facilitated by the Administrative Data Service (ADS) who will work with the data owners to ensure that a legal gateway exists for the access.

Access to the data via the VML is only possible through controlled rooms, located in ONS and other Government buildings. Before access is made available at any given location, it has to be approved by ONS.

Researchers will take no data into the room, and will take out no data. Results from the research will only be released by ONS staff once they have confirmed that they contain no risk of identifying an individual, household or business.

7.2 Physical Security Measures

The servers used to store data, and to host the analysis programs are located within a Pan-Government Accredited (PGA) data centre, based in the UK. This data centre has a comprehensive range of physical security measures including:

- access control via card and pin
- CCTV
- security guards
- intruder alarms
- independent checks of both physical and electronic security processes carried out by accredited UK security organisations.

Access to the rooms storing these servers is strictly controlled, and only data centre staff with appropriate security clearance have unaccompanied access. No other access is allowed, other than by appointment and with an escort. Researchers never have access to this room.

The terminals used by researchers to access data only provide a “window” to the data. All processing is completed on the secure servers, with no data, code or analysis copied onto the local terminal at any time.

7.3 Digital Security Measures

In addition to protecting the servers hosting the VML, additional steps are also taken to ensure security of data during transfer and analysis.

The VML is only accessible via the Public Sector Network (PSN), a secure Government network, never from the public internet.

When it is necessary to transfer data between the VML and ONS servers, to deposit or remove data, this can only be done in one of two secure ways:

- By transferring encrypted files via the PSN
- ONS staff visiting the VML server room who transfer the data using a secure CESG-approved encrypted media.

All aspects of the data storage and analysis environment are subject to real-time protective monitoring, to ensure that any unauthorised attempt to access the data is identified and acted upon immediately by the security manager. ONS have the ability to immediately block ALL access to data stored within the VML if considered necessary.

8 Conclusion

In conclusion, we have put in place processes and procedures to ensure that privacy and security are fully protected throughout all stages of ONS processing of data for ADRN projects. These processes may change as we refine our techniques, but any changes will not compromise the security and privacy safeguards set out in this paper.

Appendix A: Operating procedures of the linkage facility

Administrative data are collected for a specific administrative purpose such as processing a benefit claim or recording a patient's change of address. The collection and recording of the data follow a certain process which helps to meet the requirement. These data may also be of use for statistical purposes although this is not the main reason for collecting the data. The data therefore need to be prepared to allow the maximum statistical benefit to be gained from them. The processes laid out in this paper are in large part the same as the processes researched and defined during the initial phase of the Beyond 2011 Programme,¹¹ They have been subject to an Independent Methodological Review and accepted as fit for purpose.

Linkage is the process of joining two datasets where each dataset comprises a series of records where each record contains details of an individual. We join together the records using a set of identifiers (usually name, address, and date of birth, although others can also be used). Once we know which records from the different datasets refer to the same individual, their attributes can then be joined together. It is the attributes that the researchers need in order to pursue their research and the identifiers are only needed to identify which records relate to the same person. One of the ADRN principles is that no-one working in the ADRN (linkers or researchers) should have access to identifiers and attributes of people in the admin datasets, and we achieve this by deleting the identifiers prior to bringing in the attributes.

The linkage process outlined here specifically will not join individuals to households, or households to geographic areas. Both of these are possible and likely to be required, but will require different techniques to be developed and are not considered in this paper.

Once the linkage has been achieved in the most effective manner possible, we then carry out an assessment to ensure that the resulting dataset does not impact on the privacy of the individuals whose data are included (noting that no direct identifiers will be released from the linkage facility to the researchers). We also ensure that the secure environment has sufficient safeguards and controls to handle the specific dataset.

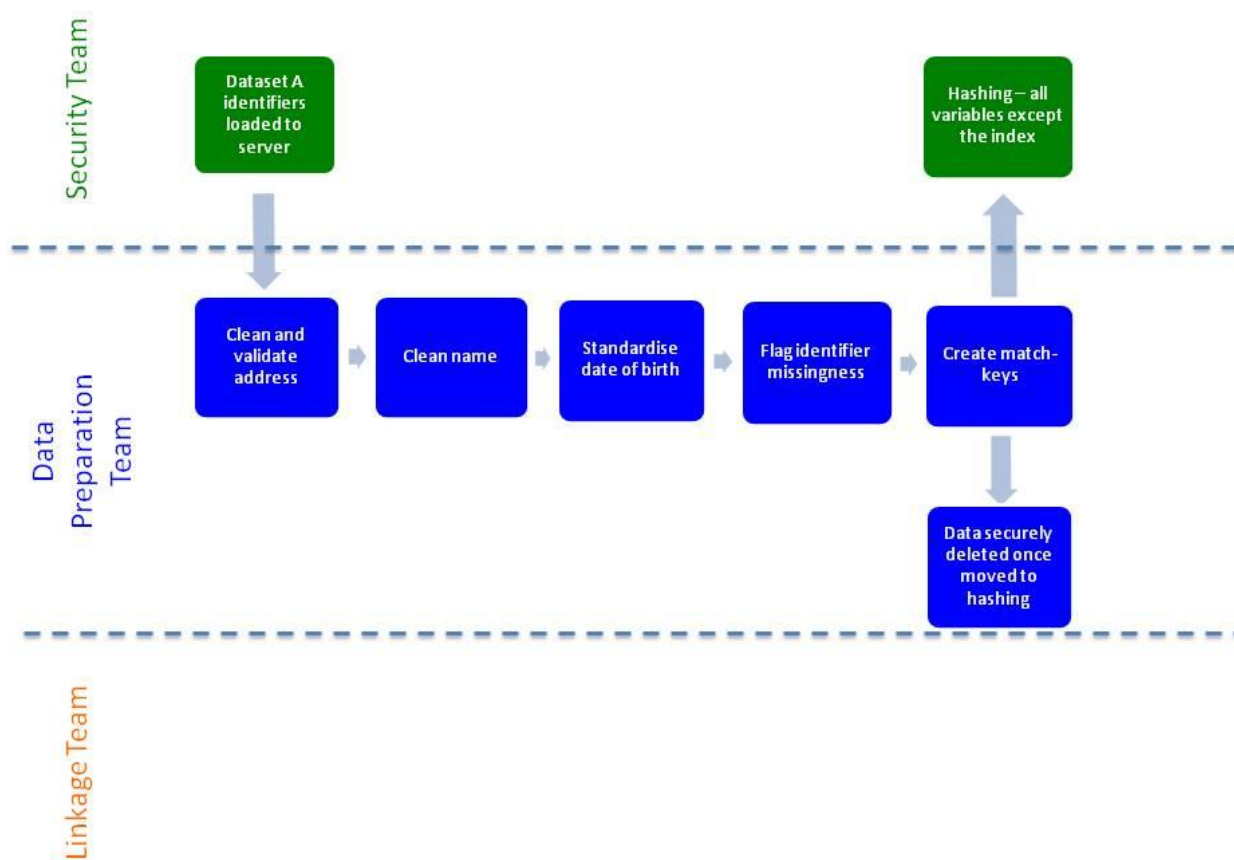
A.1 Data Preparation

This section outlines the process through which each separate dataset passes in order to prepare the data for linking to other datasets. It describes the journey from raw administrative data to administrative data that can be linked.

The administrative data that ONS receive are in the format used by the administrative system and often arrive with many different formats. In order to link together two or more datasets, it is important that they have the same format otherwise valid links may be missed. For example, in two different datasets, someone born on the first of January 2001 might be presented as: 01/01/2001, 1/1/2001, or 1/1/01 1st Jan 2001, 01-Jan – 01 etc. We need to move all of these to a common format so that they can be linked.

The diagram below shows the process that the dataset of identifiers goes through before being linked to the identifiers of the second dataset.

¹¹ For more on the B2011 programme: <http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/index.html>



This section provides detailed explanations of how certain variables are cleaned and the reasons why.

1.1 Reasons for cleaning

Statistical data are well organised and exhibit clean, tidy characteristics that make them fit for purpose. Administrative data, whilst fit for administrative purposes may be held in many different formats across organisations and there are many different people responsible for compiling and maintaining them. This means that data cleaning is required for the data to be fit for statistical use.

Administrative data are therefore highly likely to contain variations in the identifiable variables which make them difficult to match. Cleaning can help to standardise these variations and therefore improve the chances of linking records between datasets. Cleaning will not help resolve errors in the data – for example if Jon Smith noted his name as Jom Smith, then cleaning will not help.

In the following sections we will consider various challenges that are found in the data and how we go about addressing them in order to create useable statistical data.

1.2 Challenges we face

Administrative datasets exist in various different formats, SAS, CSV (comma separated variable), Microsoft Excel etc. Variables come in various formats, character, numeric, several date formats

and text. The same variables could also occur in different field lengths between datasets. All these factors need to be cleaned and standardised to maximise linking potential.

Even where data are electronically collected, there can be variations in the format of the information. For example, one dataset might have forenames in different variable fields, forename 1, forename 2, etc. Another dataset might have all forenames grouped into one forename field.

Unexpected characters occur in different quantities between datasets and are unpredictable in terms of where they occur. It is usually the result of an administrative system where variables are filled in free format text fields and therefore prone to keying error.

1.3 Specific items of code cleaning performed

1.3.1 Address field cleaning

Address fields in administrative data sometimes contain characters that would not normally be expected to form part of an address. This can provide an issue with validating postcodes between datasets as different sources might contain different characters which could result in different or failed validation.

Any characters other than alphanumeric, apostrophe, hyphen or ampersand are removed from the records and replaced with a space.

All datasets are passed through sophisticated geo-referencing software which matches all the address information provided against various reference files and returns a validated current postcode in standardised format.

1.3.2 Name field cleaning

In a similar way to the address cleaning, Name fields can appear with characters other than letters. This can halt processing and lead to missed matches.

Name fields will be cleaned by replacing characters other than English letters, apostrophe or hyphen, with a space. You could expect to see these characters in some names but any other characters would be considered spurious and are removed.

Names can also be quite complex and not easy to fit into a standard forename, middle-name and surname format. Data sometimes come with more than one name in the forename or middle-name fields. The processing needs to try and standardise this by splitting these cases out into new fields such as forename 1 and 2, middle-names 1 and 2. In the case of forename, this is done by searching for the first instance of a space in the forename field and then putting the string before the space into forename 1 and the string after the space into forename 2. This increases the chance of making matches between datasets.

In order to prepare for the matching process, the aim is to produce a dataset with forename, middle-name and surname variables populated. This will enable all of the match-keys, which are described in more detail later in this paper, to be created.

1.3.3 Handling Duplicate Records

Data providers provide data with a unique random index attached which ensures that there are one-to-one links between identifier and attribute files. This unique index will be checked on both the identifier and attribute datasets. Any instances of non-uniqueness will be referred to the data provider as the matching process requires each record in the data to refer to a unique individual.

The matching process includes safeguards against duplication. It can only make a match between datasets A and B where the match-key in question is unique on both sources of data. If this is not the case and duplicate matches are identified, they will not be output as a match.

1.3.4 Values

If variables are missing then it could create problems in the matching process. In particular, matches could be made which may not necessarily exist if a two records were considered to match for a specific variable just because the value was missing in both datasets.

We will therefore establish, with the help of the data provider, which variables are missing prior to the matching process. We then insert standard code to ensure that the missing variables in source A are coded as '88888888' and source B '99999999', for example.

This will ensure that the records cannot match on the missing variable and therefore acts as a further control to ensure reliable matches.

1.3.5 Standardising the “Sex” Variable

Datasets from different sources will contain a few variations for defining the sex variable. For example, the records for a male could be denoted by a sex variable of Male, M or 1, with the Female being F or 2.

Whichever format in which this variable appears in the dataset, it is coded to the same value across all datasets to ensure accurate comparison. The standard coding is as follows;

1 – Male, 2 – Female, 3 – Other, 4 – Not stated

1.4 Match-keys

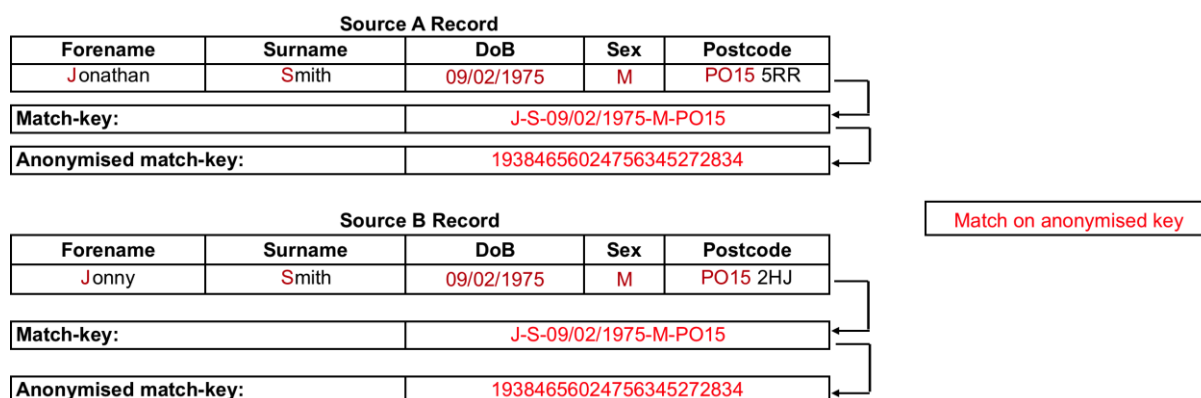
The match-keys are variables created within our linkage process which are consistent across all datasets to enable the matching process to take place.

Match-keys are created by putting together pieces of information to create unique keys that can be used for automated matching; the intention is the elimination of some of the discrepancies that might otherwise prevent an automated match. For example, a match-key can be constructed from the initials of an individual's forename and surname, combined with their date of birth, sex and postcode district¹². The resulting string is then hashed¹³ and can be used to link records between datasets in the linkage part of the environment.

¹² Postcode District is the inward part of the postcode e.g. SW19

¹³ A form of anonymisation, see section 2 for details.

Figure B.1: Example of match-key creation for linking records



Our approach is to construct a series of these match-keys, each of which is designed to resolve a particular type of inconsistency that often occurs between records belonging to the same individual. These match-keys are presented as matching fields in a hierarchical, or stepwise, linkage process, each forming a separate 'match pass' essentially forming the deterministic phase of the overall matching strategy¹⁴

A major requirement when using this approach is to ensure that the resulting match-key retains a high level of uniqueness for the majority of records to be matched. An example of a match-key produced in this way is a concatenation of forename initial, surname initial, sex, date of birth and postcode district. Having undertaken frequency analysis of the Patient Register (PR), it is estimated that 99.55 % of people in the UK have a unique match-key when data is concatenated in this way. Furthermore, the potential for disagreement between two matching records is significantly reduced when matching on this information - provided the first letter of the forename and surname have been documented correctly and the individual has accurately reported the characters of their postcode district, the potential for inconsistency between the two sources has been significantly reduced.

Inconsistency between matching variables can occur in a number of different forms. A single match-key alone cannot resolve all of the inconsistencies that occur between data sources. Frequency analysis of the PR has been undertaken for a range of variable concatenations resulting in a series of match-keys which are being used in our matching approach, all of which are designed to resolve particular inconsistencies between match pairs. Figure 2 presents the structure of each of these match-keys and the uniqueness of those keys when they are created for all records held on the de-duplicated 2011 PR. It also summarises the type of inconsistency that each match-key is designed to resolve. These match-keys capture, on average, 95 % of the available matches. This has been calculated by comparing to the matches made by a 'gold standard' matching approach utilising clerical resolution in an un-hashed environment.

¹⁴ Such hierarchical deterministic approaches are prevalent in linkage studies across epidemiology (e.g. Li *et al.*, 2006; Pacheco *et al.*, 2008) and the match-key approach has been seen to perform well in an Australian community health care study (Karmel *et al.*, 2010).

Figure B.2: Uniqueness of match-keys derived from the de-duplicated 2011 PR and the inconsistencies they resolve between true match pairs

Match-key	Unique records	Inconsistencies resolved by match-key
(1) Forename, Surname, DoB, Sex, Postcode	99.99%	None - exact agreement
(2) Forename, Surname, DoB, Sex	98.90%	Movers out of area
(3) Forename Initial, Surname Initial, DoB, Sex, Postcode District	99.55%	Name / postcode discrepancies
(4) Forename Initial, DoB, Sex, Postcode	99.89%	Surname discrepancy
(5) Surname Initial, DoB, Sex, Postcode	99.47%	Forename discrepancy
(6) Forename, Surname, Sex, Postcode	99.21%	DoB missing / incorrect
(7) Forename bi-gram ² , Surname bi-gram, DoB, Sex, Postcode Area	99.45%	Name discrepancies / movers in area
(8) Forename, Surname, Year of Birth, Sex, Postcode District	99.47%	DoB discrepancy / movers in area
(9) First Middle Name, Surname, DoB, Sex, Postcode	99.86%	Forename / middle name transpositions
(10) Second Middle Name, Surname, DoB, Sex, Postcode	99.53%	Forename / middle name transpositions
(11) Forename, Surname, DoB, Postcode	99.99%	Sex missing / incorrect

² In this context the bi-gram comprises the first two characters of the name

Data sources containing alternative name or location for a record can be accommodated in the process. It requires a separate set of match-keys to be created in pre-processing, which are then matched with the other source to maximise potential matches. This does involve defining which is the 'preferred' match-key set, so that only residuals from the preferred match are put through the alternate match. This has worked well, for example, with the Higher Education Student Authority (HESA) data, which records a term-time and domicile (typically parent or guardian) address. The term-time address is used in preference when linking to PR data, as that is where they reside, but in the event that they haven't registered with a doctor at their university town, it is likely they are on the PR at their domicile address. We therefore have a better chance of matching them to the PR with the domicile address match-keys.

The 11 Match-keys¹⁵ that we currently use are as follows¹⁶:

Match-key 1	FORENAME – SURNAME – DOB – SEX – POSTCODE
Match-key 2	FORENAME – SURNAME – DOB – SEX
Match-key 3	FORENAME INITIAL – SURNAME INITIAL – DOB – SEX – POSTCODE DISTRICT
Match-key 4	FORENAME INITIAL – DOB – SEX – POSTCODE
Match-key 5	SURNAME INITIAL – DOB – SEX – POSTCODE
Match-key 6	FORENAME – SURNAME – SEX – POSTCODE
Match-key 7	FORENAME BIGRAM – SURNAME BIGRAM – DOB – SEX – POSTCODE AREA

¹⁵ All identifiable data fields in all data sources being used to create the linked dataset are pseudonymised. In line with the Information Commissioner's Code for Practice for Anonymisation http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation this means that all uniquely identifiable fields within each dataset are cryptographically hashed, allowing them to be electronically matched across datasets in such a way that the original identifiers are not available to those doing the linkage.

¹⁶ Note, these Match-keys can be changed depending on the data to be matched.

Match-key 8	FORENAME – SURNAME – YEAR OF BIRTH – SEX – POSTCODE DISTRICT
Match-key 9	1 ST MIDDLE-NAME – SURNAME – DOB – SEX – POSTCODE
Match-key 10	2 ND MIDDLE-NAME – SURNAME – DOB – SEX – POSTCODE
Match-key 11	FORENAME – SURNAME – DOB – POSTCODE

Therefore the match-keys for Joe David Jason Bloggs born on 1st Jan 1999, Male with postcode of PO15 5RR would be as follows;

Match-key 1	JOE – BLOGGS – 01/01/1999 – MALE – PO15 5RR
Match-key 2	JOE – BLOGGS – 01/01/1999 – MALE
Match-key 3	J – B – 01/01/1999 – MALE – PO15
Match-key 4	J – 01/01/1999 – MALE – PO15 5RR
Match-key 5	B – 01/01/1999 – MALE – PO15 5RR
Match-key 6	JOE – BLOGGS – MALE – PO15 5RR
Match-key 7	JO – BL – 01/01/1999 – MALE – PO
Match-key 8	JOE – BLOGS – 1999 – MALE – PO15
Match-key 9	DAVID – BLOGGS – 01/01/1999 – MALE – PO15 5RR
Match-key 10	JASON – BLOGGS – 01/01/1999 – MALE – PO15 5RR
Match-key 11	JOE – BLOGGS – 01/01/1999 – PO15 5RR

1.5 Hashing

After cleaning and matchkey production, the processed data are hashed in order to make sure the records cannot be identified and in this process each identical record receives the same coded hash value. This enables identical records to be matched across various datasets. If there are erroneous characters in records, then identical hash variables will not be created and matches will be missed.

It is important to note that different keys will be used for different matching projects. This means that while, for example, 'John' will be hashed to the same value on both sources within a project, it would have a different hashed value in a different project.

In order to comply with legal, regulatory and inter-departmental security requirements, as well as to ensure individual privacy is safeguarded, a strong cryptographic hash¹⁷ is used to ensure uniquely-identifying variables within source data are not intelligible to those undertaking the linkage in line with the Information Commissioner's Office Code of Practice relating to anonymisation¹⁸.

ONS has intentionally selected proven industry standards to allow public scrutiny of our approach – the only aspect that needs to be kept "secret" is the set of hashing keys used in operation. This is quite technical and an overview has been provided separately in appendix C.

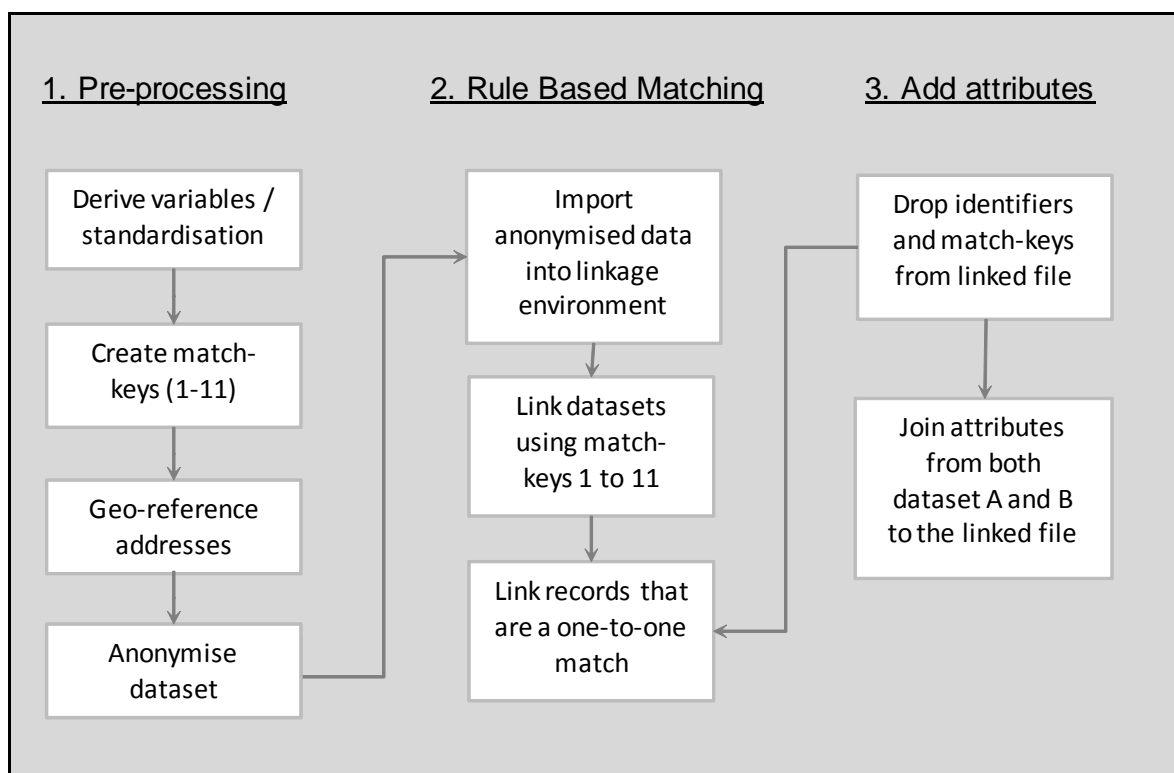
¹⁷ Note that "hashing" is a one-way, irreversible process, as opposed to "encryption", which is designed to be reversible

¹⁸ "Anonymisation: managing data protection risk code of practice" - http://www.ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation - our approach aligns with "Case Study 11".

A.2 Linking

The two datasets are joined to each other by each of the match-keys in turn, in order of uniqueness according to the analysis done for the PR 2011. Records are only linked on a match-key if it is unique on both datasets (i.e. one-to-one match). If multiple records match on a particular match-key then the link is not made and candidates are passed on as a residual to the next match pass. The hierarchical nature of the whole matching process has implications. Matches that are made at an early stage of the process are linked and removed, with only the residuals being passed to the next stage.

Figure B.3: ADRC-E matching strategy for the SRE



Once the unique matches have been created the linkage identifiers generated during the pre-processing are used to link the attribute data.

2.1 Outputs

As variables such as date of birth and postcode are identifiers they cannot be included in the linked dataset; however age at a given point in time and geographies such as MSOA can be added if needed. Where the same variable appears on both datasets, such as age or location, the requester would nominate a 'master' dataset so the variables from this dataset are used should the two sources differ.

Any unlinked records from both datasets feed into summary statistics and provided as an unmatched output. Metadata on the matching process will also be provided.

2.2 Future developments in linking methods

The set of match-keys can be developed and extended with future research, such as one based on match-key five (Surname Initial, DoB, Sex, Postcode) but including the whole surname rather than just the initial. This would be more discriminatory than the current match-key five but would still allow the forename to be spelled incorrectly and potentially make a match.

Currently, each match-key is only matched to their respective one on the other source. This could be developed to allow match-key nine, which replaces forename with middle name, to be matched to match-key one on the other source, thus capturing forename and middle name transpositions. This could be extended to forename and surname transpositions with the creation of a new match-key: Surname, Forename, DoB, Sex, Postcode, which would be matched to match-key one on the other source.

There are ways of using logistic regression to make further matches from the residuals. This involves creating similarity tables using string comparisons prior to hashing each source. Any development in this area will need to balance the accuracy of the linkage, the quality and quantity of the data and the resource needed to identify additional linkages.

A.3 Pre-release Assessment Process

Once the two datasets containing attributes have been joined together in the ONS linkage facility and prior to releasing the linked data to the secure environment to be used for research, the newly formed linked and de-identified data is owned by the Director for the Administrative Data Division. One of the responsibilities of the Director is to conduct a pre-release assessment which will ensure that:

- The privacy of the individual is maintained, and
- the designated secure environment has appropriate security measures in place to securely manage the dataset.

If the assessment process indicates concern over either privacy or security then these will be addressed prior to the dataset leaving the linkage facility¹⁹.

3.1 Privacy considerations

The linked datasets created will not contain any personal identifiers; there will be no names, addresses, dates of birth or references such as National Insurance Number. They will however contain information which could increase the chances of someone with additional knowledge being able to identify someone. For example people reaching 100 often receive coverage in the local press and a researcher may therefore be able to 'recognise' someone with similar characteristics, in terms of age, gender and location, in the linked dataset.

We will carry out additional checks on the linked data to minimise the chance of identification using the information present. These checks will review the tables that can be constructed using the main identifying variables in the dataset to ensure there are at least three people in each category. The types of identifying variables which will be considered are age, gender, ethnicity and household size. This approach addresses some of the concerns referred to in the Ipsos MORI

¹⁹ It should be noted that the approvals process for ADRN projects considers the sensitivity of the required linked datasets during discussions with the data owners. This should minimise the risk of subsequent problems arising, although it is impossible to be sure of data sensitivities until the two or more sources have been linked.

[‘Dialogue on Data’](#). The ADRC-E data scientists and researcher(s) will be informed of exactly which variables will be checked prior to the linkage work starting.

For example there may be fewer than three males over the age of 100 living in some areas, this means the data cannot be released with an upper age category of 100+ at that level of geography and with a split between gender. In this case the data could be made available if a 90+ category were used. Alternatively the data could be released using a higher level of geographic referencing or without the gender split. In all cases the approach taken to address privacy concerns will depend upon the underlying research question. Any remedial action will be discussed with the data scientists and researcher(s) before it is implemented, although the need for security and privacy is of paramount importance. Researcher preferences help where there are different choices that can lead to safe data.

The following guidelines provide some assistance when specifying the linked dataset to be created, although many other considerations will be needed for each specific case.

1. The lowest level of geographic referencing which will be available on a linked dataset is Lower Layer Super Output Area (LSOA). LSOAs have an average population of 1,500 people as at the 2011 Census.
2. Unmatched records will only be made available if the approved research proposal relates to those cases. If that is the case then the matched records will not be released²⁰.
3. Due to the size of the populations it may not always be possible to provide data referenced to the Isles of Scilly and the City of London. In these cases the Isles of Scilly will be combined with Cornwall and the City of London with Westminster.

3.2 Security of data within the Secure Environment

The current approach for defining the security measures required by a dataset is through the use of CESG [Business Impact Levels](#)²¹. The approach considers the impact of a breach of confidentiality (a data loss), integrity (data corruptions) and accessibility (data unavailability) to assign a Business Impact Level (BIL). A BIL is assigned for all Government datasets, including those provided to the Linkage Facility to be linked for ADRC-E projects, and the linked de-identified dataset that the researcher will access in the secure environment.

Each secure environment will have provided sufficient assurance to accept data up to a given BIL. We will only release data to a secure environment if its accreditation permits holding data with the BIL assigned to the linked de-identified dataset.

In assessing Business Impact Level, the assessor firstly considers which of the main BIL tables is/are relevant: Defence, International Relation, Security and Intelligence, Public Order, Public Safety and Law Enforcement, Trade, Economics and Public Finance, Public Services, Critical National Infrastructure (CNI), Personal/Citizen.

For the majority of ADRC-E projects the Personal/Citizen table will be the most relevant, although if business related administrative data are involved, then Trade Economics and Public Finance may also need to be considered.

Within the table for Personal/Citizen there are a number of sub-categories that need to be considered. These are: Impact on health and safety of the Citizen, Impact on the Privacy of the

²⁰ Both data owners and the Approvals Panel would need to approve cases where linked and unlinked data are provided to the researchers. This may be possible for non-personal or small sample data.

²¹ Communications-Electronics Security Group, the National Technical Authority for Information Assurance.

Citizen, Impact on the Identity of the Citizen. Utilisation of Public Services, Embarrassment or Distress, Personal Finance

As an example, the following table provides Guidance on Impact Levels for Privacy of the Citizen.

BIL Level	Guidance on impact
BIL 0	None
BIL 1	Loss of control of a citizen's personal data beyond those authorised by the citizen
BIL 2	Loss of control of many citizen's personal data beyond that authorised by each citizen
BIL 3	Loss of control of a citizen's sensitive data beyond those authorised by the citizen. A compromise to the identity or financial status of an individual citizen.
BIL 4	Loss of control of many citizens' sensitive or financially significant personal data beyond those authorised by each citizen. A compromise to the identity or financial status of many citizens, increased vulnerability to criminal attack.
BIL 5	Widespread compromise of identity management systems or personal financial systems across the UK.
BIL 6	The collapse of identity management systems or personal financial systems across the UK

As can be seen, as the Impact Level rises from 0 to 6, so do the consequences from None to a collapse of UK wide systems.

The Assessor is given the BIL assessments for the input datasets. These were produced by the owner of the data. For a single record (individual) in the dataset from an administrative source, it is likely that the impact will be at least BIL1 or 3 depending on whether there are sensitive data in the dataset.

If the two input datasets for linking have different BIL assessments, then the linked dataset will, as a minimum, take on the higher BIL level.

In the pre-release assessment discussed here, the only consideration is whether joining the data together has increased BIL from the higher of the input datasets. Questions such as: "do the combination of characteristics lead to additional sensitivity of the data? Might the combination of characteristics lead to a potential compromise of identity management systems?"

Other broader questions that lead to the assessment of BIL of the linked dataset are:

For each input dataset what population does it cover?

- What is the geographic coverage?
- What population does it cover and are there any known exclusions such as armed forces?
- Are the sources administrative data or survey data?
- What is the reference period of the data? Is the data as at a particular point in time or has it been accumulated over a period of time?

How would you describe the matched records and the unmatched records?

- Other than data quality issues what real life circumstances could lead to a record being present in one data source and not in the other?

Can population subgroups be identified?

- Do the data include information that could be used to infer membership of a particular subgroup? For example pensions being taken before normal retirement age or locations close to known military establishments.

Is any of the information in the dataset commercially sensitive?

- Do the data include information about sole traders or partnerships?

Does this data contain information that could put an individual's safety or security at risk?

- Can you identify particular accommodation types such as sheltered housing or domestic abuse hostels?
- Do the data allow the tracking of individuals over time?
- Could the data lead to identity theft or a loss of privacy?
- If the data were compromised how much distress and embarrassment could it cause an individual?

The Assessor will look across all the tables and sub-categories and provide answers to such questions in a report. This will be reviewed by the owner of the linked dataset whilst it is stored in ONS (the Divisional Director of the Admin Data Division), who will authorise it.

If the level of risk associated with a dataset is above the threshold for the designated secure environment the researcher may have the option of carrying out their work in an alternative location that is accredited to the appropriate level. If this is not possible then, as with privacy concerns, then the options are to apply remedial action to the dataset with a view to reducing the impact level, or declaring that the research, as originally planned, cannot proceed.

Remedial action may involve recoding some variables. Advice will be sought from the data owners as to the most suitable approach if this is required. Any remedial action will be discussed with the researcher prior to being implemented.

The data owner will be informed if the BIL associated with the actual linked datasets is different from that presumed when they gave permission for their data to be used. In this case they will be asked to confirm that they are still content for the research project to proceed in the proposed secure environment.

In this situation a decision requested whether to proceed with the research as this will constitute a change in the processing instructions that the data owner originally gave the Linkage Facility.

Appendix B: Detailed specification for the cryptographic functions in use

In order to address the privacy and security concerns that may arise when administrative datasets are linked in the ONS secure linkage facility, we have implemented a defence-in-depth approach to protection of the administrative source data being processed. While this reduces the need for robust defence against cryptographic attacks (due to other technical, physical and procedural controls in place), the approach is designed to resist all currently-feasible attacks on the basis that the same approach may be used in future (and/or elsewhere) in less controlled environments.

It must be noted that this “pseudonymisation”²² approach is **not** designed to eliminate the risk of Statistical Disclosure, for which additional governance and processes are in place (see section 3 of this appendix), it is solely focused on protecting individual-level personally-identifiable information. The approach consists of two stages:

Stage	Description	Regularity
1	Per-field Key Generation For each field that needs to be hashed within source datasets, and for each derived “matchkey”, a unique, random key is generated and held securely for as long as required.	Once only per linkage project – any change will require all source data to be reprocessed.
2	Hashing Operation For every linkage project a source dataset is imported, each identifying variable and “matchkey” is hashed using a key-based hashing algorithm (HMAC), which combines the field content with the unique key from Stage 1 to produce a final consistent, irreversible value based on the original field input.	As often as required – once per field each time a source dataset is imported.

Notes

1. By design, in order to maximise utility, the concepts described herein are platform- and language-agnostic – where possible, references to public Standards are provided, all of which should be available in industry-standard cryptographic libraries, and implemented in the majority of common programming languages.
2. In order to assure “safe” application of the described cryptographic operations, well-tested cryptographic libraries are used, and all code is independently reviewed to ensure correct use of input and output parameters (including checks for correct “casting” or conversion of variables between string, byte, binary and any other types).

²² The ICO’s Code of Practice describes the difference between “anonymisation” and “pseudonymisation” – the hashing technique described is more strictly described as “pseudonymisation”.

Stage 1 – Per-field Key Generation

In order to defend against some forms of theoretical attack, random keys are generated for use in the HMAC process in Stage 2. One unique random key is generated per field that needs to be matched, with the same key being used for similar fields across all datasets within a project– e.g. a “Postcode Key”, a “Forename Key”, a “Matchkey 1 Key”, etc.

Ideally, these keys should be truly random, or at least generated using a Cryptographically Strong Pseudorandom Number Generator (CSPRNG), but this can be difficult/expensive to implement, depending on the available hardware/software platform. The ADRC-E Linkage Facility takes the CSPRNG approach, but provided the keys are handled appropriately, the entropy in the salt is not a critical factor in the final hash strength, and thus deterministic pseudorandom number generators could be used – most programming languages have some form of function available for this.

The length of the keys are 1024 bits (128 bytes, or 256 hexadecimal digits), as this is the optimum key size for the HMAC-SHA512 algorithm used in Stage 2.

Example Key (for Test Output Validation)

A hexadecimal representation of a 1024-bit key would look similar to the following (on a single line – word-wrapping unavoidable in this document):

```
FF37CE1D748EE255BA057E3727234E8467A899E1C1836132E6EEC6C9B80AE6987D4D8F784
224120BFFE5B877287616A9B769FE20469728A77168C2254D6ABC626F169221EFE8E0AAC2
78F147DA8BC960E25951C93A9E1CA4C924CB91E6F2227621036EA2E064E6D23F84CB02B4F
2E9AC3568A447FFB33503B06F48156CF345CD
```

Stage 2 – Hashing Operation

The hashing algorithm selected for the ADRC-E Linkage Facility purposes is the Hash-based Message Authentication Code (HMAC), using SHA-512 as the hashing algorithm (HMAC-SHA512). RFC2104²³ details the HMAC process (including the static “opad” and “ipad” parameters), but as detailed earlier the process uses calls to cryptographic libraries to ensure no errors are introduced – the parameters defined in RFC2104’s pseudocode are as follows:

H (Hashing Algorithm)	SHA-512
K (Key)	The per-field key derived in Stage 1
text	Raw value of the field to be hashed
ipad/opad	Static values, as detailed in the RFC

The final output from this process is **always** a 512-bit (64-byte) hash value, which is represented as a hexadecimal string (128 characters).

Since the use of multiple 128-character strings in operation can lead to storage and performance issues, truncation of this value has been used in order to reduce overhead – the first 128 bits (16 bytes, 32 hex digits) of the hash value are used in the final datasets. This level of truncation does not introduce a significant risk of collision²⁴, whereas further truncation might.

²³ <http://www.ietf.org/rfc/rfc2104.txt>

²⁴ “Collision” is where multiple different raw values produce an identical hash, resulting in a reduction in statistical quality and confidence.

Example Input/Output (for Test Cases)

The following table provides examples of the expected format for input and output parameters at this Stage. Again, hexadecimal string representations have been word-wrapped.

Attribute	Description	Example(s)
K	The keys produced in Stage 2 for each field to be hashed in the input dataset.	7722265BC678B1D8977542F517DCCC16 964BF5999A600C394AE6B1C7543A3F91 4A8792C31B6833EFAB18DCF62E5F1A99 31D90B9B7764D861B8C310EFED4BA02E 52014D312DDFC3B79F1A6028E968F37A D360CE64BF2BFEE3C56D954AD87CB5E7 BC3E420AAB0F9C86A445BBB153A6BDED AEAB67E62E66B52E40A9118EE465A318
text	The text contents of the field to be hashed.	"John" "Smith" "1982-05-17" "ZY99 2XA"
Full Output	The final (full) hash value for the field concerned.	9CF9BC9AD0240662E90E459E193E38DF 22FCABEA45357C1B30C0877C7168AF44 FFE9493E94FBD2B76FBFFC1E2874E1F1 E9151279350C2AEDE4FEF46484744D95
128-bit Truncation	The first 16 bytes (32 hex digits) of the final output hash.	9CF9BC9AD0240662E90E459E193E38DF

Example Hash Values

The following table gives an example of what the identifiable data fields, such as name, address and date of birth, look like once they have been hashed. The examples here demonstrate that variables with the same original value will be hashed to the same final value (for matching purposes), and that once pseudonymised, the data bears no resemblance to the original value.

Please Note: The information set out below is for illustrative purposes only and does not contain any information relating to a real person. Non-operational test keys have been used in generating these examples.

String/Value to Hash	Hashed Value
John	8C 17 A3 BB 4C AF 71 9D 16 50 97 90 0B 39 01 61
Smith	39 E9 3E D6 6E 50 A7 EC 6B F9 4F 9B 9F CF 81 F6
Jon	86 1A 42 1C 1A 05 E0 E8 FA 24 A1 53 41 59 69 1F
Smith	39 E9 3E D6 6E 50 A7 EC 6B F9 4F 9B 9F CF 81 F6
John	8C 17 A3 BB 4C AF 71 9D 16 50 97 90 0B 39 01 61
Smyth	CB 36 9F C9 0A 3B A0 2E E9 9C A0 5E E0 69 84 FB
Jonathan	F4 5C C5 B7 A6 59 23 79 B8 5B 81 81 AA AD 38 50
Smith	39 E9 3E D6 6E 50 A7 EC 6B F9 4F 9B 9F CF 81 F6
Jonny	ED ED 5C 0E 56 00 83 84 AA 03 8F E7 02 AA AB E3
27/01/1965	4F 6E B0 E4 55 84 BC 0A 8B A3 89 B5 16 F4 49 9A
26/01/1965	2C 5A 2C 3D 80 D1 48 35 70 24 6A D8 E5 2C 94 17
27/01/1965	4F 6E B0 E4 55 84 BC 0A 8B A3 89 B5 16 F4 49 9A
27/02/1965	EC 67 CC 6D C7 23 40 84 09 E7 B5 7C DE 79 6B D4

27/01/1966	90 BF F8 D3 C5 DD 3F DB 3C 6D DC 39 AD EE E6 46
1965	10 B2 57 F8 08 7E 72 F1 2A E6 96 E4 A1 E4 26 DE
1966	5C C9 4A 4C CA E8 48 75 B5 52 68 E0 B0 C5 F3 CA
1965	10 B2 57 F8 08 7E 72 F1 2A E6 96 E4 A1 E4 26 DE
1966	5C C9 4A 4C CA E8 48 75 B5 52 68 E0 B0 C5 F3 CA

Example: Name and date strings transformed into hashed values

String to hash	Hashed value
John	8C 17 A3 BB 4C AF 71 9D 16 50 97 90 0B 39 01 61
Smith	39 E9 3E D6 6E 50 A7 EC 6B F9 4F 9B 9F CF 81 F6
Jon	86 1A 42 1C 1A 05 E0 E8 FA 24 A1 53 41 59 69 1F
Smith	39 E9 3E D6 6E 50 A7 EC 6B F9 4F 9B 9F CF 81 F6
John	8C 17 A3 BB 4C AF 71 9D 16 50 97 90 0B 39 01 61
Smyth	CB 36 9F C9 0A 3B A0 2E E9 9C A0 5E E0 69 84 FB
Jonathan	F4 5C C5 B7 A6 59 23 79 B8 5B 81 81 AA AD 38 50
Jonny	ED ED 5C 0E 56 00 83 84 AA 03 8F E7 02 AA AB E3
27/01/1965	4F 6E B0 E4 55 84 BC 0A 8B A3 89 B5 16 F4 49 9A
26/01/1965	2C 5A 2C 3D 80 D1 48 35 70 24 6A D8 E5 2C 94 17
27/01/1965	4F 6E B0 E4 55 84 BC 0A 8B A3 89 B5 16 F4 49 9A
27/02/1965	EC 67 CC 6D C7 23 40 84 09 E7 B5 7C DE 79 6B D4
27/01/1966	90 BF F8 D3 C5 DD 3F DB 3C 6D DC 39 AD EE E6 46
1965	10 B2 57 F8 08 7E 72 F1 2A E6 96 E4 A1 E4 26 DE
1966	5C C9 4A 4C CA E8 48 75 B5 52 68 E0 B0 C5 F3 CA

Once the data have been hashed, the identifiers are deleted as they have served their purpose. It is the hashed match-keys that move to the next stage in the process where records relating to the same person are joined together.

Appendix C: Glossary of Terms used

ADRC-E. The Administrative Data Research Centre for England	Led by the University of Southampton, and run in collaboration with: University College London, the London School of Hygiene and Tropical Medicine, the Institute for Fiscal Studies, the Office for National Statistics.
ADRN - Administrative Data Research Network	The Administrative Data Research Network (ADRN) is a UK-wide partnership between academia, government departments and agencies, national statistical authorities, funders and the wider research community that will make it easier to carry out economic and social research based on routinely collected government administrative data.
ADS - The Administrative Data Service	The Administrative Data Service (ADS) coordinates the Administrative Data Research Network, and is the first point of contact for researchers who want access to administrative data. It is based at the University of Essex within the UK Data Archive, with partners at the Universities of Edinburgh, Manchester, Oxford and the West of England.
Anonymisation – de-identification	The process of removing direct identifiers from datasets so that they can be safely used for research purposes.
Approvals Panel	Panel that assesses if a project can be granted access to de-identified administrative data ensuring that the approval process is fair, equitable and transparent. The AP will review project proposals based on ethics review, privacy impact assessment, feasibility, scientific merit and public benefit.
Business Impact Level	A government measure of the sensitivity of data used to accredit information handling.
CESG	Communications Electronics Security Group is a branch of the Government Communications Headquarters working to secure the communication and information systems of the government. CESG is the UK National Technical Authority for Information assurance
Data Controller	The person or organisation defined by the Data Protection Act as being responsible for determining the purpose for which and the manner in which a particular dataset containing personal data is processed.
Developed Vetting	UK National Security Vetting for individuals employed in posts requiring them to have long-term, frequent and uncontrolled access to Top Secret assets.
Linking and Matching	The process of joining together two different administrative datasets. Linking refers to the process of taking identifier data (such as name and address) to determine whether the two different datasets refer to the same person. Matching is the process of joining together the attributes from the different datasets having first determined that they are the same person.
Linkage Facility	The computers and procedures that permit secure linking of administrative datasets within ONS.
Pan Government Accreditor	The person appointed by HM Government to accredit systems that span more than one government department or which lie outside government departments but which access government data.
Principal Investigator	The lead researcher for a specific project. This individual is responsible for defining the project, and for determining the team of researchers who will work on the project.
Protective Marking	Applied to an asset to indicate the sensitivity, or confidentiality, and thus the level of protection required.
PSN. Public Sector	A public Sector secure network for sharing information. Used by the VML

Network	to administer projects.
Secure Environment	Access point for researchers to carry out their safe research on de-identified data. The researcher can access statistical programmes to carry out their research, but cannot remove any data or outputs
Security Check	UK National Security Vetting for individuals employed in posts requiring them to have long-term, frequent and uncontrolled access to Secret assets.
Statistical Disclosure Control	Statistical techniques for assessing and ensuring that outputs are safe and cannot be used to reveal confidential information
VML. Virtual Microdata Laboratory	The ONS flagship laboratory for providing safe access to de-identified ONS data for research purposes