Administrative Data Research Network

An ESRC Data Investment

ADRN GLOSSARY

Definitions of Terms used in the Network

EXTERNAL

21 NOVEMBER 2016

Version: 00.09

T +44 (0)1206 873435
E help@adrn.ac.uk
www.adrn.uk

Scope

This document is a Glossary for (technical) terms used in the ADRN. The Glossary should contain, but not be limited to, all terms used in ADRN documentation.

Purpose

The definitions outlined in this document are intended to be of guidance in the harmonisation process of the procedural documents produced within the Network.

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

Administrative Data

Data about individuals collected by government departments and agencies for operational purposes in the day-to-day delivery of government services.

Administrative Data Research Centre (ADRC)

An *Administrative Data Research Centre (ADRC)* is a consortium of academic institutions with expertise in looking after data, using *Administrative Data* for research and making them available for other researchers. The *ADRN* has four *ADRCs*, based in England, Northern Ireland, Scotland and Wales.

Administrative Data Research Network (ADRN)

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 new economic and social research based on routinely collected government *Administrative Data*.

Also referred to as 'The Network'.

Administrative Data Service (ADS)

The *ADS* coordinates the *Administrative Data Research Network*, and is the first point of contact for researchers who want access to *Administrative Data*.

ADRN Board

The independent governing body of the *ADRN*, chaired by the UK Statistics Authority. The *ADRN Board* reports directly to the UK Parliament on the *ADRN* to ensure the robust performance and governance of the Network. For further details on its main roles and the *ADRN Board* members, please see the UK Statistics Authority website.[1]

ADRN Directors Group

The *Directors Group* manages the Network and consists of:

- the directors of the four *Administrative Data Research Centres* and the *Administrative Data Service*
- the *ESRC* lead for data and resources
- a representative from the governing board
- the Chair of the *Operations Group*

1 http://www.statisticsauthority.gov.uk/national-statistician/administrative-data-research-network/index.html

The *Directors Group* reports to the Economic and Social Research Council[2], which is accountable to the Department of Business, Energy & Industrial Strategy[3] (formerly Department of Business, Innovation and Skills).

The *Directors Group* is responsible for all Network policy, strategic forward planning, resource allocation and innovation.

ADRN Operations Group

The *Operations Group* manages the Network day to day, and is responsible for collective decision-making on how to implement the strategic direction decided by the *Directors Group*.

The *Operations Group* consists of two representatives from each *ADRC* and two from the *Administrative Data Service*. The representatives are appointed by each unit's Director, and one of them must be that unit's Project Manager.

The *ESRC's* Senior Policy Manager will also attend. Others, such as the Chair of a task team, may be invited when relevant.

ADRN Researcher

A researcher, from academia, the public sector or a research organisation on the Research Councils UK *(RCUK)* list of eligible *Independent Research Organisations (IRO)* [4] that:

(i) is a trusted 'fit and proper' person i.e. they must be capable of carrying out the research either independently or under the direction of an appropriate supervisor or lead investigator;

(ii) has been granted access to *De-identified data* by the respective *Data Controller*(s) on an *ADRN* research project;

(iii) has successfully completed ADRN Safe User of Research data Environments (SURE) training;

(iv) is backed by an *Institutional Guarantor*, and

(v) has signed up to the ADRN Terms of Use (ADRN021)[5].

2 http://www.esrc.ac.uk/

3 https://www.gov.uk/government/organisations/department-for-business-energy-and-industrial-strategy

4 http://www.rcuk.ac.uk/documents/documents/eligibilityiros-pdf/

5 https://adrn.ac.uk/media/1161/adrn014-ap_termsofreference_01_02_pub.pdf

ADRN Safe User of Research data Environments (SURE) Training

The training that ADRN researchers are required to complete before they can access any data provided through the ADRN to ensure that they are able to access data safely, securely and lawfully.

For training to be considered complete, the researcher must attend a SURE training session, actively engage with the training session and pass a test.

Anonymisation

A process of rendering *Data* into a form which does not identify individuals and where identification is not likely to take place. The process can use various techniques, such as the coding or masking of directly, or indirectly, identifying *Data*.

Anonymous Data

Data whereby, even when combined with other sources of information in the public domain, the identity of *Data Subject*s cannot be determined. *Anonymous Data* are typically available to the public (i.e. open data).

Approvals Panel (AP)

The Panel that assesses whether a project can be granted access to de-identified *Administrative Data* and makes sure that the approval process is fair, equitable and transparent. See ADRN014 – Approval Panel Terms of Reference[6] and ADRN015 – Approval Panel Operating Procedure.[7]

Approved Institution

An organisation from academia, the Public Sector or a Third Sector organisation on the Research Councils UK list of eligible Independent Research Organisations (IROs)[8] determined as part of the eligibility criteria for using ADRN services.

Also referred to as "Institution".

Approved Researcher

For *ADRN* purposes, the preferred term is *ADRN Researcher*.

Please note that the term Approved Researcher is defined differently in the Statistics and Registration Services Act 2007.[9]

6 https://adrn.ac.uk/media/1161/adrn014-ap_termsofreference_01_02_pub.pdf

7 https://adrn.ac.uk/media/1163/adrn015-ap_operatingprocedure_01_01_pub.pdf

8 http://www.rcuk.ac.uk/funding/eligibilityforrcs/

9 http://www.legislation.gov.uk/ukpga/2007/18/pdfs/ukpga_20070018_en.pdf

Artificial Data

Data that have been generated randomly to provide the structure of a *Dataset* without representing actual *Data* values.

Attribute Data

Data that are used for research purposes. These are the data (with *Direct Identifiers* removed) provided by the *Data Controller* to one or more of the *ADRC*s or the Office of National Statistic's secure *Data Linkage* facility.

B

Breach

A *Breach* can refer to a *Data Breach* or a *Security Incident*.

Breaches Policy

The *ADRN* Controlled Document [ADRN003][10] which outlines *ADRN* policy and procedures for managing breaches of the ADRN Terms of Use [ADRN021][11], and other *ADRN* security procedures, by *ADRN Researchers*.

The *ADRN* is committed to protecting the privacy and confidentiality of respondents while promoting good research practice. This document supports these key commitments by providing a framework for applying penalties for misuse of *Sensitive Data* and breaches of *ADRN* procedures. The penalties outlined in the document do not exclude the possibility of criminal penalties.

C

Content Data

Preferred term: *Payload Data.*

10 https://adrn.ac.uk/media/1297/adrn003_breachespolicy_02_pub.pdf

11 https://adrn.ac.uk/media/1399/adrn021-termsofuse_v00-10_pub.pdf

D

Data Breach

Incident in which *Data* is unlawfully removed from a *Secure Environment.*

A *Personal Data* breach means "a *Breach* of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised *Disclosure* of, or access to, *Personal Data* transmitted, stored or otherwise processed in connection with the provisions of a public electronic communications service".[12]

Data Cleansing

Data Cleansing is the process by which *Data* are prepared for research. This process usually entails editing, labelling, *Metadata* creation and checking for anomalies.

Data Collection

Data Collection can be used as a synonym for *Dataset*, but includes *Dataset*(s), documentation and *Metadata*.

Data Controller

A 'person' recognised in law, i.e. individuals, organisations, other corporate and unincorporated bodies of persons, who (either alone or jointly or in common with other persons) determines the purposes for which, and the manner in which, any *Personal Data* are, or are to be, processed.[13]

Data Custodian

This term is used on the *ADRN* website to address a public audience but refers to the *ADRN* preferred term: *Data Controller*.

Data Linkage

The process by which records about a single *Data Subject* across multiple *Datasets* are associated with each other to create a new 'linked' *Dataset*.

Data Matching

The process by which *Direct Identifiers* from different data sources are used to identify common *Data Subjects*. Alternatively, multiple sets of *Direct Identifiers* can be matched against a population spine (eg a count by age, sex and small area) to identify common *Data Subjects*.

Within the *ADRN Data Linkage* process, the output of the *Data Matching* process is typically a set of *Matched IDs* that can be used by the *Linker* to link the *Research Data*.

12 Information Commissioner's Office (ICO) – Security Breaches: https://ico.org.uk/for-organisations/guide-to-pecr/security-breaches/

13 Data Protection Act 1998 – definition:http://www.legislation.gov.uk/ukpga/1998/29/section/1/enacted

Data Owner

The legal definition of this term is: A legal entity that authorises or denies access to data for which it maintains accuracy, integrity and timeliness.

For *ADRN* purposes, the preferred term is *Data Controller* and is used as a more precise definition.

Data Processor

This term, in relation to *Personal Data*, means any person (other than an employee of the *Data Controller*) who processes the data on behalf of the *Data Controller*.[14]

Data Protection Act 1998 (DPA)

The piece of UK legislation which regulates the processing of information relating to individuals, including the obtaining, holding, use or *Disclosure* of such information.[15]

Data Provider

Preferred term: *Data Controller*.

Dataset

A *Quantitative Dataset* is a collection of structured information on *Data Subject*s that can be numerically measured and statistically analysed obtained using quantitative research methods such as surveys or questionnaires.

A *Qualitative Dataset* is a collection of unstructured information on *Data Subjects* that typically cannot be numerically measured obtained using qualitative research methods such as interview transcripts, audio/video/digital recordings and photographic material.

Data Subjects

*Data Subject*s are the individuals, households or organizations on which information (i.e. *Data*) is collected.

De-identification

The process used to prevent an individual's identity from being connected with information that relates to them. Common strategies for de-identifying *Dataset*s are deleting or masking personal identifiers, so *Direct Identifiers* such as name, postcode and date of birth, might be concatenated and replaced with a unique reference.

De-identified Data

Extracts from *Data* which have undergone the process of *De-identification*.

14 ICO definition: https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/

15 Data Protection Act 1998: http://www.legislation.gov.uk/ukpga/1998/29/contents/29/contents

Direct Identifiers

Direct Identifiers refers to variables (or sets of variables) in *Datasets*, such as name, address, full date of birth, postcode information telephone number and tax reference number, which can directly identify subjects.

In connection with the *Data Linkage* process, *Direct Identifiers* refer to data extracts containing only *Direct Identifiers* and a *Project Specific Unique Identifier*.

Disclosure

The act of releasing identifying information to unauthorised third parties.

Disclosure Control

Techniques used to control the risk of individuals being identified from *Statistical Data*. Typical methods include removing or disguising *Data* relating to individuals with unusual sets of attributes.

Preferred terms: *Statistical Disclosure Control (input) - Statistical Disclosure Control (output).*

E

ESRC

The Economic and Social Research Council.[16]

Ethical Review

A review of the ethical implications of a project and how these issues are addressed. The main aim of an *Ethical Review* is, as far as possible, to protect all groups involved in research: participants, institutions, funders and researchers, throughout the lifetime of the research and into the dissemination process.[17]

F

Feasibility Report

The report, generally produced by *ADS* and *ADRC*(s) staff, to assess the feasibility of an *ADRN* project in terms of demonstrating a need for the *ADRN* service and the *ADRN* being able to support the research and provide a safe setting. The *Feasibility Report* is part of the official documentation submitted to the *Approvals Panel* for all *ADRN* project applications. This assessment of feasibility should be considered indicative rather than definitive, as the assessment is made at a relatively early stage in the project preparation process.

16 http://www.esrc.ac.uk/http://www.esrc.ac.uk/

17 ESRC Framework for Research Ethics: http://www.esrc.ac.uk/files/funding/guidance-for-applicants/esrc-framework-for-research-ethics-2015/

File Transfer

The secure transfer of *Data* across the network. This includes transfer via an encrypted media, or through an encrypted internet/network connection.

I

Independent Research Organisation (IROs)

An organisation eligible to apply for and receive research funding by the Research Councils UK (*RCUK*). They usually "possess an existing in-house capacity to carry out research that materially extends and enhances the UK research base, and are able to demonstrate an independent capability to undertake and lead research programmes".[18]

Indexer

Preferred term: *Matcher*.

Indexing

Preferred term: *Data Matching*.

Index Key

Preferred term: *Project Specific Unique Identifier*.

Indirect Identifiers

Indirect Identifiers refers to variables (or sets of variables) in *Datasets*, such as information on workplace, occupation or exceptional values of characteristics like salary or age, which, when linked with other publicly available information sources, could identify subjects.

Input SDC

See *Statistical Disclosure Control (input)*

Institutional Guarantor

An individual, within an *ADRN Researcher*'s institution, with the legal status to act on behalf of that institution (e.g. the Director of Research Grants and Contracts, or equivalent post). The *Institutional Guarantor*[19] is responsible for endorsing *ADRN Researcher* applications and accepting liability, on behalf of the institution, for any serious and/or consistent breaches of security by those researchers, in relation to ADRN Breaches Policy (ADRN003).[20]

18 Research Councils UK: http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/eligibilitystatement.pdf

19 https://adrn.ac.uk/media/1158/adrn012-institutionalguarantorpolicyandprocedure_01_01_pub.pdf

20 https://adrn.ac.uk/media/1297/adrn003_breachespolicy_02_pub.pdf

L

Lead Researcher

For ADRN purposes, the single point of contact for progress updates and communication regarding the progress of the application.

Legal Gateway

The laws which allow for access of *Personal Data*.

Linked Data

A *Dataset* that is created through *Data Linkage*.

Linker

Staff at an *ADRC* facility who perform the *Data Linkage*.

Note: The *Data Linkage* process is divided into *Data Matching* and *Data Linkage*.

Linking

Preferred term: *Data Linkage*.

M

Matched IDs

Typically, two or more sets of *Project Specific Unique Identifier*s and their relation, identifying common *Data Subject*s in the source *Dataset*s.

Matcher

Staff at a *TTP* who perform the *Data Matching*. In the case of an automated *Data Matching* process, *Matcher* refers to staff managing this process.

Metadata

Metadata are information that helps a researcher understand the *Data*. Typically, *Metadata* includes information about data type, data collector, time period, geographical coverage and sampling procedures.

Microdata

Microdata are unit-level data, i.e. *Data* on individual *Data Subject*s rather than on aggregates of larger groups of *Data Subject*s.

O

Output Checking

Preferred term: *Statistical Disclosure Control (output).*

Outputs

Research outputs that have been released by *ADRC* after *Statistical Disclosure Control* (*output*).

Intermediate outputs may be released during the analysis, for the purpose of improving the research. These outputs can only be shared with accredited researchers in the team.

Final outputs are released for publication purposes.

Output SDC

See *Statistical Disclosure Control (output).*

P

Payload Data

Data that are used for research purposes. These are the data (with *Direct Identifiers* removed) provided by the *Data Controller* to one or more of the *ADRC*s.

Personal Data

Personal Data means data which relate to a living individual who can be identified:

(a) from those *Data* or

(b) from those *Data* and other information which is in the possession of, or is likely to come into the possession of, the *Data Controller*,

and includes any expression of opinion about the individual and any indication of the intentions of the *Data Controller* or any other person in respect of the individual. Where the ability to identify an individual depends partly on the *Data* held and partly on other information (not necessarily *Data*), the *Data* held will still be *"Personal Data"*.[21]

Data are considered *Personal Data* if a living individual can be identified from the *Data*, or, from the *Data* in combination with other information (publicly) available.[22]

Personal Information

Information that relates to and identifies an individual (including a body corporate) taking into account the other information derived from any other published sources (as defined in clause 39 of the Statistics and Registration Service Act 2007).[23]

21 https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/

22 http://www.legislation.gov.uk/ukpga/1998/29/section/1

23 http://www.legislation.gov.uk/ukpga/2007/18/section/39

Primary Data

Data that were previously unknown and which have been obtained directly by the researcher for a particular research project.

Principal Investigator

The lead researcher of a research project.

Privacy Impact Assessment (PIA)

The assessment of the potential impact of an *ADRN* research project on individuals' privacy. The assessment determines the possible risk and probability of *Disclosure*, including the mitigation of any residual risks. The *PIA* is part of the official documentation submitted to the *Approvals Panel* (*AP*) for all project applications. All Privacy Reports are countersigned by *ADS.*

Project Proposal

Research outlined in the ADRN Project Proposal form, processed by the Network, that needs to be submitted by a *User* to request the *Data Linkage* of, and/or access to, *Administrative Data* through the *ADRN*. The information the proposal provides helps to assess whether *ADRN* is a suitable source of *Data* for the project and forms the basis of any application to the *Approvals Panel (AP)*.

Project Specific Unique Identifier

In the process of separating *Direct Identifiers* from *Research Data*, a *Project Specific Unique Identifier* is typically produced. This is a number uniquely identifying a *Data Subject* in each *Dataset* without disclosing any additional information.

Pseudonymised Data

Preferred term: *De-identified Data*.

Public Engagement

Public Engagement is a two-way process, involving interacting with and listening to the public, with the goal of generating mutual benefit.

Q

Qualitative Data

Qualitative data are unstructured information on *Data Subject*s collected using research methods, such as participant observation or case studies to record people's attitudes, feelings and behaviours in greater depth, which result in a narrative, descriptive account of a setting or practice which typically cannot be numerically measured.

Quantitative Data

Quantitative data are structured information on *Data Subjects* collected using research methods, such as surveys or questionnaires which allow for the measurement of variables, within a collection of people or groups, and resulting in numerical data which can be subjected to statistical analysis.

R

Research Councils UK (RCUK)

Research Councils UK (RCUK)[24] is the strategic partnership of the UK's seven Research Councils. Each year the Research Councils invest in research covering the full spectrum of academic disciplines from the medical and biological sciences to astronomy, physics, chemistry and engineering, social sciences, economics, environmental sciences and the arts and humanities.

Research Data

Preferred term: *Payload Data*.

In connection with the *Data Linkage* process, *Research Data* refers to data extracts strictly containing no *Direct Identifiers* but typically including a *Project Specific Unique Identifier*.

Researcher Application

The application form which needs to be completed by each researcher intending to participate in an *ADRN* project. The form includes details about themselves, their research and experience with accessing *Sensitive Data*. This information helps to assess whether they are eligible to become an *ADRN Researcher* as part of an *ADRN* approved project. The completed form is submitted to *ADS* via the stated *Institutional Guarantor*.

S

Safe Data

Preferred term: *De-identified Data*.

Data handled within the *ADRN* are not referred to as safe or secure. The *Secure Environment* used to access the data provides one of the key safeguards and assurances that data can be accessed safely and securely.

Safe Environment

Preferred term: *Secure Environment*.

24 http://www.rcuk.ac.uk

Safe Haven

Preferred term: *Secure Environment*.

Safe Pod

A pre-fabricated and demountable *Secure Room*. Safe pods offer additional security and assurance over using *Secure Remote Access* from institutional desktops, though use the same information technology to connect to secure server rooms.

Safe Room

Preferred term: *Secure Room.*

Safe Setting

Preferred term: *Secure Environment.*

SDC

See *Statistical Disclosure Control (input), Statistical Disclosure Control (output).*

Secondary Data

Data that has already been collected for a purpose other than the current research project which may also have value for research.

Secure Access

Accessing *De-identified Data* within a *Secure Environment*.

Secure Data

Preferred term: *De-identified Data*.

Data handled within the *ADRN* are not safe or secure. The *Secure Environment*, used to access the *Data*, provides one of the key safeguards and assurances that *Data* can be accessed as safely and securely as possible.

Secure Data Centre

A facility where *Data* can be stored and processed securely. Access to the data centre, and the servers is controlled through appropriate physical, technical and procedural controls.

Processing *Data* held on the server is typically carried out from a *Secure Room* or *Safe Pod*.

Secure Environment

A physical or virtual facility where an *ADRN Researcher* can access the administrative *Data* requested in their approved project proposal. This facility will be hosted at one of the *ADRC*s or indicated by the organisation holding the *Data*. *Secure environment* refers to a suite of *Secure Access* infrastructures, including but not limited to *Secure Server Room, Secure Room, Safe Pod, Secure Remote Access* solutions.

Secure Remote Access

The method by which *Data* on a secure server can be remotely accessed and processed in a secure manner, typically from specific institutional desktops. Access to secure remote access solutions is controlled through appropriate physical, technical and procedural security measures.

Secure Room

Dedicated room used to access and process *De-identified Data*. The *Data* is typically stored in a *Secure Server Room*, though stand-alone secure rooms exist, in which case the *Data* is held locally in the room. Access to machines in the room, and the room itself is controlled through appropriate physical, technical and procedural security measures.

Secure Server

A Web server which uses data encryption and decryption protocols to protect *Data* from authorised interception.

Secure Server Room

A secure area in which the *Secure Server* is located, designed to specific requirements to ensure the security, maintenance and monitoring of the *Secure Server.*

Secure Setting

Preferred term: *Secure Environment*.

Security Incident

An incident in which security measures (physical, technical or procedural) are not followed or are breached.

An information security incident is a single or series of unwanted or unexpected information security events that have a significant probability of compromising business operations and threatening information security. An information security event refers to an identified occurrence of a system, service or network state indicating a possible *Breach* of information security policy or failure of safeguards, or a previously unknown situation that may be security relevant.

Sensitive Data

Sensitive *Personal Data* consisting of information as to a *Data Subject*'s:

(a) racial or ethnic origin
(b) political opinions
(c) religious beliefs or other beliefs of a similar nature
(d) whether they are a member of a trade union (within the meaning of the Trade Union and Labour Relations (Consolidation) Act 1992[25]
(e) physical or mental health or condition
(f) sexual life
(g) the commission or alleged commission by him of an offence, or
(h) any proceedings for any offence committed, or alleged to have been committed by him, the disposal of such proceedings or the sentence of any court in such proceedings.

Staging Environment

The processing environment where *Payload Data* is combined using *Matched IDs*. This environment is used to create derived variables (where applicable) and where, prior to *Data Linkage*, *Metadata* on *Unlinked Data* can be derived (where applicable). The staging environment is not accessible by researchers.

Stakeholder

A person, group or organisation with an interest or concern in an organisation. Stakeholders can affect or be affected by the organisation's actions, objectives and policies.

Statistical Data

Information which is held in the form of numerical data, nominal data (eg gender, ethnicity, region), ordinal data (age group, qualification level), interval data (month of birth) or ratio data (age in months).

Statistical Disclosure Control (input)

Methodology used in the design of statistical products in order to protect the identity of *Data Subject*s.

Statistical Disclosure Control (output)

Methodology used to check that research outputs do not disclose any identifying information before release from a *Secure Environment*.

25 http://www.legislation.gov.uk/ukpga/1992/52/contents

Survey Data

Survey data are data collected through surveys, typically for the purpose of producing statistics. By participating in a survey, *Data Subject*s may give general consent to using the data collected for research purposes.

Syntax

Programming code, in a language specific to a software package that is developed by ADRC staff and Accredited Researchers in processing data for research purposes.

Synthetic Data

Data that have been generated to reflect actual data without containing any identifying information, or information relating to individuals.

T

Trusted Third Party (TTP)

A Trusted Third Party (TTP) performs the matching of *Direct Identifiers* from different data sources, or the matching of *Direct Identifiers* of a single data source against an existing population spline.

Within the ADRN linkage process, a TTP typically produces a set of *Matched IDs* that can be used by the *Linker* to link the *Research Data*

A TTP can be based at a *Data Controller*, or at an *ADRC*, in which case there needs to be a clear separation of roles between *Matchers, Linkers* and *ADRN Researchers*.

U

Unique Identifier

Preferred term: *Project Specific Unique Identifier*.

Unlinked Data

Attribute Data associated with *Data Subject*s that have not been matched across *Dataset*s.

User

User refers to any person who submits a request for information or support which requires action from *ADRN* User Service staff. The term also includes *Users* of any *ADRN* service which includes, but is not limited to, eligible *Accredited Researcher*s and applicants.

V

Virtual Safe Setting

Preferred term: *Secure Remote Access*.