Administrative Data | Agricultural Research Collection (AD|ARC)

Agricultural Research Collection – Wales (ARCW)

An Introductory Guide for Researchers

Version 2.1

08/10/2024

Authors

Dr Paul Caskie (Agri-Food and Biosciences Institute)

Sian Morrison-Rees (Swansea University/ADR Wales)

Nicholas Webster (Welsh Government/ADR Wales)

Laura Madden (Welsh Government/ADR Wales)

Sarah Lowe (Welsh Government/ADR Wales)

On behalf of the wider AD|ARC team in Wales: Sean Scully (Swansea University), Jenny Thyer (Swansea University), Katy Addison (Welsh Government/ADR Wales), Liam Crowley (Welsh Government/ADR Wales), Matthew Curds (Welsh Government/ADR Wales), Sorcha Egan (Welsh Government/ADR Wales), Matthew Kelly (Welsh Government/ADR Wales), Rosie Kirk (Welsh Government/ADR Wales), Rachael Loftus (Welsh Government/ADR Wales), Richard McFerran (Welsh Government/ADR Wales), Stuart Neil (Welsh Government), and Freya Pryce (Welsh Government/ADR Wales).

Funding

This work is supported by ADR UK (Administrative Data Research UK), an Economic and Social Research Council (part of UK Research and Innovation) programme and by the Welsh Government.

Acknowledgements

The AD|ARC project in Wales notes with thanks the many collaborators who have supported the development of the Agricultural Research Collection – Wales (ARCW), including the Welsh Government, Public Health Wales, the SAIL Databank, and the AD|ARC Scientific Advisory Board and Wales Stakeholder Reference Group.

Similarly, the AD|ARC team notes with thanks the support of other data linking projects in the development this introductory guidance, most notably the Data First and ECHILD projects.

This report describes the Agricultural Research Collection – Wales (ARCW), which includes data from Welsh Government, Rural Payments Wales and the ONS. These parties do not accept any responsibility for any inferences or conclusions derived by the authors.

Glossary of Terms and Acronyms

ALF	Anonymised Linking Field – an individual-level anonymised field based on a double-encrypted version of that individual's NHS Number that allows anonymised data about that individual to be linked in a secure environment for research purposes
Beta Testing	Beta testing is the process of giving access to a dataset to a small subset of potential users to identify any problems before allowing wider access.
Blocking	Blocking is when you split a dataset into 'blocks' or clusters of records using a particular attribute. This limits the number of potential matches according to the blocking attribute, improving efficiency while maintaining accuracy. For instance, blocking on place of birth and sex means that a linking algorithm looking for Angharad Rees born in Ceredigion would only search within the set of candidate matches of women born in that county.
CBUA	Contiguous Built Up Area classification – this classification is used by AD|ARC as a proxy for rurality
Cleaning	Data cleaning is the process of addressing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data in a dataset. Cleaning methods vary depending on the variable or dataset but may be as simple as fixing a clear typo to deleting invalid data altogether.
CPH	County Parish Holding
CRN	Rural Payments Wales Customer Reference Number
Data Dictionary	A Data Dictionary is a set of information describing the contents, format and structure of a dataset or database
Deterministic Matching	Deterministic Matching uses the exact correspondence between instances of an identifier or set of identifiers to produce a match. Deterministic matching assumes that unique identifiers such as names or dates of birth can be used to definitively link across datasets. Also see Probabilistic Matching.
DHCW	Digital Health and Care Wales – provides the national digital and data systems underpinning health and care services in Wales, holds the Welsh Demographic Service and is the Trusted Third Party for the pseudo-anonymisation process to bring data into SAIL.
Jaro-Winkler string comparisons	The Jaro–Winkler distance gives more favourable ratings to strings that match from the beginning for a set prefix length.
LSOA	Output Areas (OAs) are the lowest level of geographical area for census statistics and were first created following the 2001 Census. Lower layer Super Output Areas (LSOAs) are made up of groups of OAs, usually four or five. They comprise between 400 and 1,200 households and have a usually resident population between 1,000 and 3,000 persons.
Matchkey	Matchkeys tend to be fixed-length, compressed strings built from a combination of the words and numbers in a name or address such that relevant variations have the same match key value.
Missingness	Missingness is the state of being missing and is usually applied to missing values, where no data value is stored within a variable for a particular case e.g. a farm business or a person living in a farming household.
Population Spine	The purpose of a population spine is act as a reference population of individuals. The unit for a population spine is usually the person or individual. Each person in the population should be uniquely identified and included only once on the population spine. Once added, a

	person should not be removed – instead, their status in the population should be changed.
Probabilistic Matching	The process of using statistical analysis to determine the overall likelihood that two records match. Probabilistic matching is the preferred method for matching large data sets when unique identifiers such as names or dates of birth are unavailable or are not of sufficient quality to be used to definitively link across datasets. Also see Deterministic Matching.
RALF	Residential Anonymised Linking Field - an address-level anonymised field that allows anonymised data about that address to be linked in a secure environment for research purposes
Research Ready Dataset / RRD	The term 'research-ready data' is used to communicate the processing of administrative data to make it available for research. However, research-readiness has multiple dimensions. In the case of AD|ARC, this means a dataset that has been built by linking a number of agricultural and wider datasets, standardising variable names and providing metadata in order to provide a resource researchers can readily access via the Trusted Third Parties of the UK.
SAIL / The SAIL Databank	ADR Wales uses the SAIL (Secure Anonymised Information Linking) Databank at Swansea University, to link and analyse anonymised data. The SAIL Databank is a DEA-accredited Trusted Research Environment.
Single Payment Scheme 2010	Single Payment Scheme (SPS or Single Farm Payment) replaced the Common Agricultural Policy (CAP) in 2003, introducing a new method of direct subsidy payments to landowners.
UPRN	The Unique Property Reference Number or UPRN is the unique identifier for every addressable location across the UK.
WDSD	The Welsh Demographic Service (WDSD) is a database of everyone registered with a GP from 1994 to the present day. Individual people who have been registered with a GP in Wales, past and present, are represented in the WDSD data as an index of unique numbers, known as the Anonymised Linking Field (ALF). The WDSD includes an anonymised residential address history – an index of numbers, one for each household in Wales, known as the Residential Anonymised Linking Field (RALF). In this way, it is possible to associate ALFs with RALFs, that is: people to homes.

Contents

1 What is the aim of this guide?

This user guide is intended to support researchers in seeking access to the Agricultural Research Collection – Wales (ARCW). It provides an overview of the Dataset, including information on data content, quality, and security.

The guide also provides guidance on the processes required to access the Collection in Wales. This includes the process of becoming an accredited researcher, guidance on how to access the SAIL Databank and the research governance processes to seek approval for access to AD|ARC for specific research projects. Links to the AD|ARC Data Dictionaries are also included.

2 What is the Agricultural Research Collection – Wales (ARCW)?

The datasets in the Agricultural Research Collection – Wales (ARCW) have been created as part of the AD|ARC project. The aim of the project is to integrate the human dimension of farming with data on farming activities, to better understand the demographic, health, education, and economic characteristics of farm households associated with different types and sizes of farm businesses. It is hoped this will provide the insight decisionmakers need to formulate better policy for the Sector, thereby enhancing the wellbeing of farmers and their households across the UK.

AD|ARC links de-identified electronic records that are already collected by departments across the governments of the UK. The AD|ARC datasets include variables from the following sources:

EU Farm Structure Survey Wales (EUFS) 2010

Rural Payments Wales (RPWD) 2010, 2013 and 2016

ONS 2011 Census Wales (CENW)

Welsh Demographic Service Dataset (WDSD)

It should be noted that the AD|ARC Project Team have not carried out any imputation on the datasets. However, imputation of various kinds may have been carried out by the original data owners.

The AD|ARC Wales cohort is defined as: **all members of households in Wales that have been successfully linked to farm businesses claiming under the Single Payment Scheme in 2010.**

The cohort includes households where no individual had reported their main occupation as ‘farming’ in the Census 2011. However, we have included these households in the cohort because they lived at an address for which a farm business received a subsidy payment under the Single Payment Scheme in 2010.

This means that the cohort contains both:

- households where at least one individual reported an agricultural main occupation in the Census 2011; and
- households where no-one reported an agricultural main occupation in the Census 2011.

Because both types of households received subsidy payments, both are considered ‘farming households’ for analysis purposes.

It should be noted that the cohort includes household members of all ages, including children.

However, it should be noted that farm businesses who chose not to claim Single Payment Scheme subsidies for which they were eligible will not appear in the RRD. Farm businesses that did not meet the Single Application rules for the Basic Payment Scheme will also not appear in the RRD. Furthermore, where there was a second household residing at a different address, such as an adult offspring who had moved but was still part of the business, they will not be included unless their address was also provided in the RPWD application.

The Agricultural Research Collection – Wales (ARCW) consists of an Individual-Level Table and a Household-Level Table, each of which is briefly described below. The methodology to produce these Tables is described in Section 3 of this User Guide.

2.1 The Individual-Level Tables

There are four individual-level tables for the AD|ARC Wales RRD – the table relating to the cohort of individuals living in ‘farming households’ ([Table Name)_INDIVIDUAL) and a table each of three control groups.

The ([Table Name]_INDIVIDUAL) includes individual-level records for all household members included in the cohort, allowing analysis to be conducted with the individual as the primary unit of analysis. The Table includes:

- the 2011 Census variables for each individual, e.g. age, gender, approximated social grade, National Statistics Socio-economic Classification, highest level of qualification, self-reported general health, whether the individual has a long-term health problem or disability etc.;
- selected agricultural variables, e.g. role on farm as provided to Rural Payments Wales, numbers of livestock etc.; and
- variables derived from the Census or agricultural sources, e.g. whether the individual reported their occupation as ‘farming’ on the Census, farm type, farm size etc.

The AD|ARC Wales Data Dictionaries include more detailed information about every variable in each dataset, including the variable type, the source of the variable e.g. Census, the format in which the data is held and the magnitude of any missingness. As noted above, the methodology to produce these Tables is described in Section 3 of this User Guide.

In addition to the records for members of households linked to farm businesses ([Table Name]_INDIVIDUAL), three further tables contain the following control groups:

- Up to 3 individuals per cohort member (i.e. per farming household member) matched on age (+/- 3 years), gender and Contiguous Built Up Area classification (CBUA) as a proxy for rurality ([Table Name]_AGEGENCBUA)
- Up to 3 individuals per cohort member (i.e. per farming household member) matched on age (+/- 3 years), gender and employment status (e.g. self-employed) in the Census 2011 ([Table Name]_AGEGENEMP)
- Up to 3 individuals per cohort member reporting their main occupation as farming (i.e. per self-reported ‘farmer’) matched on age (+/- 3 years), gender and a set of similar occupation codes in the Census 2011 ([Table Name]_AGEGENOCC)

Please note that individuals were selected for the control groups based purely on the individual characteristics listed above and not on either household size or household characteristics. So, for example, a female child living in an area with fewer than 2,000 residents may live with a lone parent in the control group but with two adults and three siblings in the cohort or vice versa. For example, this allows researchers to examine the relative household sizes of ‘farming’ households when compared with non-farming, rural controls.

For each matched control individual, the Table includes the same 2011 Census variables as for the cohort individual, however the fields for the agricultural variables are either not included or contain a null value.

The AD|ARC Wales Data Dictionary includes more detailed information about every variable in the dataset, including the source of the variables e.g. Census, the format in which the data is held, the values and the magnitude of any missingness. The methodology to produce this Table is described in Section 3 of this User Guide.

2.2 The Household-Level Table

There are two household-level tables for the AD|ARC Wales RRD – the table relating to the cohort of farming households ([Table Name]_HOUSEHOLD) and a table for a control group ([Table Name]_CONTROL_HH).

The ([Table Name]_HOUSEHOLD) includes household-level records for each farm business included in the cohort, allowing analysis to be conducted with the household as the primary unit of analysis. The Table includes:

- the 2011 Census variables that relate to the household and property e.g. the tenure of the dwelling, household size, car/van availability, central heating etc.;
- variables relating to the farm business and activity data e.g. numbers of livestock etc.; and
- variables derived from agricultural sources, e.g. farm type, farm size etc.

The AD|ARC Wales Data Dictionary includes more detailed information about every variable in the dataset, including the source of the variables e.g. Census, the format in which the data is held, the values and the magnitude of any missingness. The methodology to produce this Table is described in Section 3 of this User Guide.

There is a household level control group:

- Up to 3 households (using Household_ID) from the Census 2011 per farming household matched via the LSOA for a similar geographical area ([Table Name]_HH_Controls).

For each matched control household, the Table includes the same 2011 Census variables as for the cohort household, however the fields for the agricultural variables are either not included or contain a null value.

Please see the Data Dictionary for more information on the variables referred to above.

3 How has the data been processed to produce the Agricultural Research Collection – Wales (ARCW)?

Digital Health and Care Wales (DHCW) use the Welsh Demographic Service (WDS) data as the ’population spine’ or ‘template’ for its anonymisation process. The WDS is a database of everyone registered with a GP in Wales from 1994 to the present day. Individual people who have been registered with a GP in Wales, past and present, are represented in the WDSD data as an index of unique numbers, known as the Anonymised Linking Field (ALF). The WDSD includes an anonymised residential address history – an index of numbers, one for each household in Wales, known as the Residential Anonymised Linking Field (RALF). In this way, it is possible to associate ALFs with RALFs, that is: people to homes.

The AD|ARC RRD brings together datasets capturing information about farm households at four different levels:

1. Individual level – the individuals who make up farm households as identified by the Person_ID in the ONS Population Census 2011 and the ALF in the Welsh Demographic Survey. There may be multiple individuals in each farm household.
2. CPH level – the EU Farm Structure Survey records farm activities at a County Parish Holding level identified by a CPH number.
3. CRN level – the Rural Payments Wales data records payments made to farm businesses at a farm Business Level, identified by a Customer Reference Number (CRN). Some farm businesses in Wales have multiple County Parish Holdings.
4. Household level – the households identified by the Household_ID in the Census 2011 and the RALF in the Welsh Demographic Survey (please see information on the limitations of the RALF below in Section 4).
   N.B. There are farm households in Wales that correctly link to multiple farm businesses.

Rural Payments Wales subsidies are paid at the farm business (CRN) level. The EU Farm Structure Survey is collected at the CPH level. A farming household can have several CRNs. Equally, more than one farming household can be associated with a CRN. A CRN can have several CPHs. However, each CPH will have a distinct CRN (i.e. CPHs are not split across CRNs or Farming households).

The steps below outline how this data has been processed to create the individual and household tables mentioned above.

3.1 Preparation of the agricultural data

The linking of agricultural datasets was done within Welsh Government before the combined agricultural dataset was sent via DCHW to the SAIL Databank. The Welsh Agricultural/Farm Survey Register of addresses (used to collect the EU Farm Structure Survey) was linked to the address[1] in the Rural Payments Wales data using the following matching steps (please see Figure 1, below, for a diagram of the linking process):

a. Matched exactly on County Parish Holding (CPH) Number
b. Matched exactly on a matchkey based on ‘cleaned’ house number, postcode, and Address Fields 1 to 4 from the address in the Farm Register to the correspondence and main farmer addresses in the Rural Payments Wales data.

[1] This was either the Main Farmer address or the Correspondence Address, which Welsh Government Agricultural Statisticians advised could be, in practice, used interchangeably.

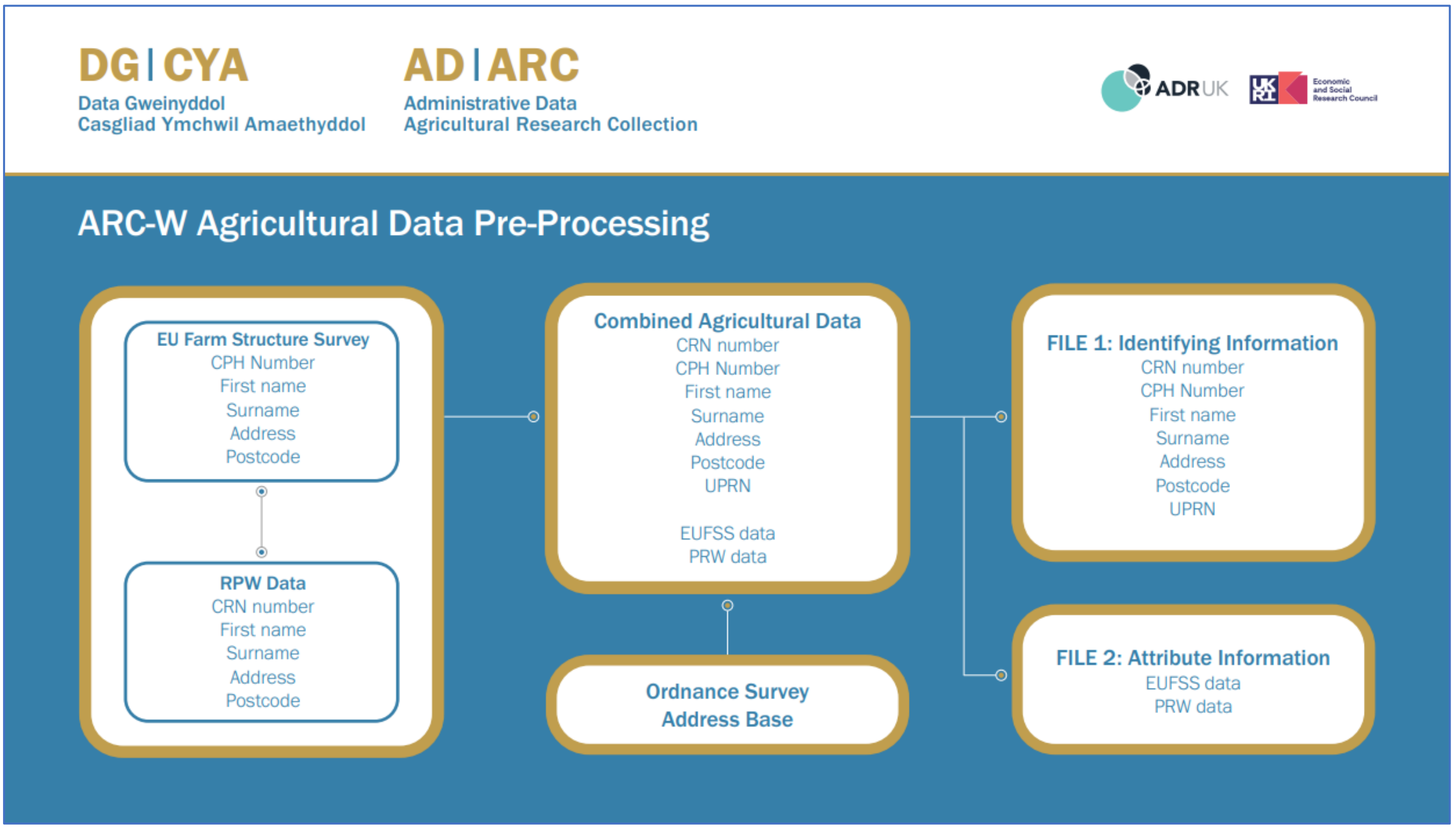
c. Matched probabilistically using Jaro-Winkler string comparisons while blocking on postcode, house number and first letter of Address Field 1.
d. If an EU Farm Structure Survey address was not matched to either RPWD address, the EUFS matchkey was used instead to probabilistically match to any of the other EUFS records and given the same CRN Number.

The Ordnance Survey Address Base Unique Property Reference Number (UPRN) was then added to the dataset to allow Digital Health and Care Wales (DHCW) to provision the data in the SAIL Databank.

A record for every farm business included in the EU Farm Structure Survey for Wales 2010 was then sent to DCHW. This contained the records for some farms no longer in operation in 2011, since there was no process for farmers to inform Welsh Government they were no longer farming. Please see Section 4 for information about linking rates.

The combined EU Farm Structure Survey-RPWD data was then uploaded to the SAIL Databank via their standard Split-File Process.

Figure 1: Diagram of ARCW Agricultural Data Pre-processing



3.2 Process of creating the cohort

As noted above, the AD|ARC Wales cohort is defined as: all members of households in Wales that have been successfully linked to farm businesses claiming under the Single Payment Scheme in 2010.

To create the cohort, one of the following two processes were followed (please see Figure 2, below, for a diagram of how the AD|ARC Datasets were linked in the SAIL Databank):

1. For combined EU Farm Structure Survey-RPWD records where, prior to de-identification, a ‘farmer’ name of useable quality was available:
    a) DCHW used the ‘farmer’ name to identify the same individual in the Welsh Demographic Survey Data – this allowed an ALF (Anonymised Linking Field) for the ‘farmer’ to be appended to the record;
    b) the de-identification was completed by DCHW and the record, now minus the ‘farmer’ name but including the ALF, was brought into the SAIL Databank;
    c) in the SAIL Databank, the ALF was used to link the record to the Census 2011, so that Census 2011 variables and record variables could be brought together for the de-identified ‘farmer’;
    N.B. some ‘farmer’ ALFs matched to more than one Census 2011 record. Where this was the case, the Census records were de-duplicated – matches who reported farming as their main occupation were accepted over matches that didn’t report farming as their main occupation. Where no matches reported farming as their main occupation, the match with the highest probabilistic matching score was accepted;
    d) the Household ID from the Census 2011 was then used to add the Census 2011 records of all of the other household members reported to be living in that residence;
    e) The farming variables for the de-identified ‘farmer’ were then appended to every member of the household so that e.g. even children’s individual records include information about farm type, livestock numbers etc.

2. For combined EU Farm Structure Survey-RPWD records where, prior to de-identification, a farmer name of useable quality was NOT available but a usable address WAS available:
    a) DCHW used the address information to match to the same address in the Welsh Demographic Survey Data – this allowed a RALF (Residential Anonymised Linking Field) for the ‘farm’ and ALFs for the ‘farm’ household to be appended to the record;
    b) the de-identification was completed by DCHW and the record, now minus the farm address but including the RALF was brought into the SAIL Databank;
    c) in the SAIL databank the ALF’s from all individuals in the Census 2011 with a stated occupation of farming and their respective households were linked to the WDSD to append the RALFs;
    d) the RALFs on the Census 2011 records were then linked to the RALFs on the agricultural data;

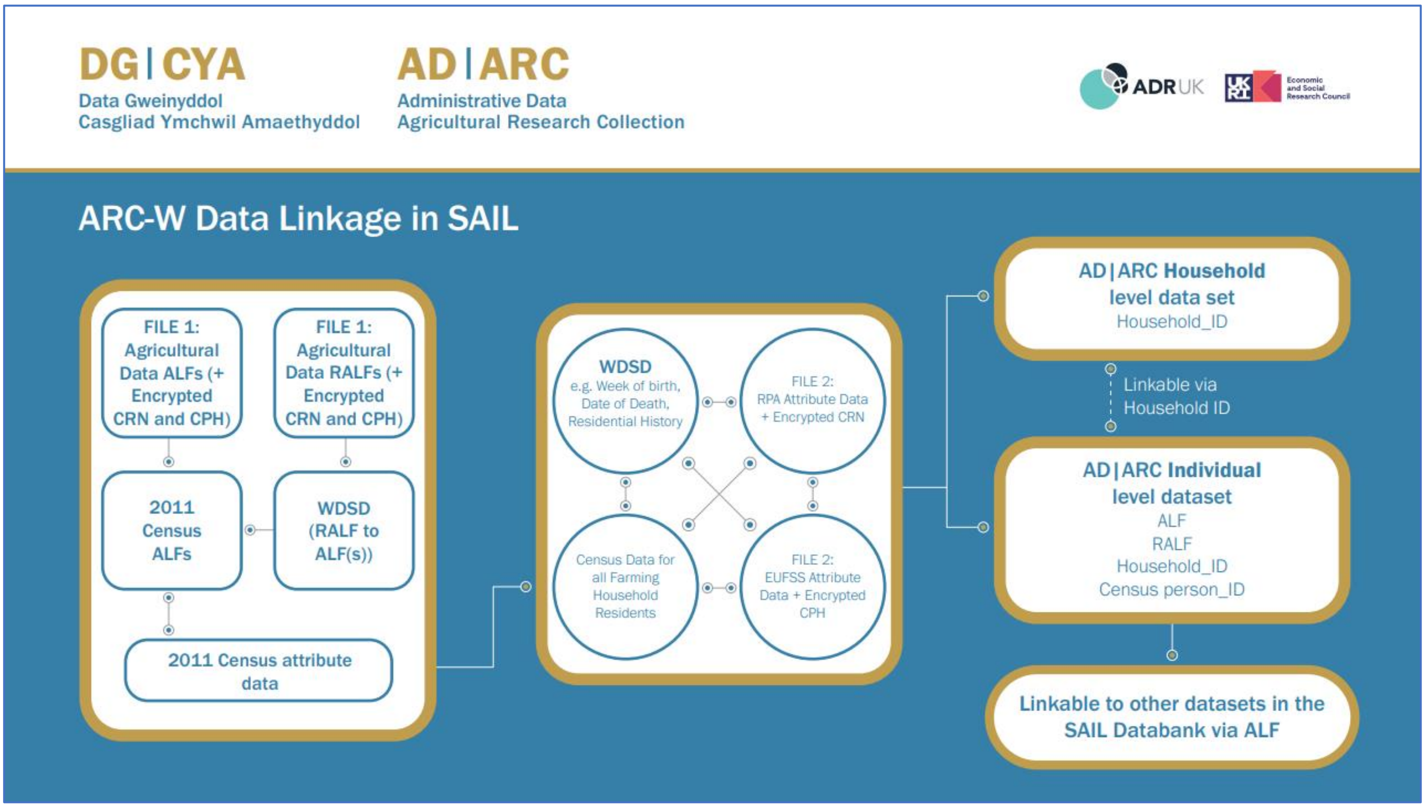
    N.B. matches via the RALF were accepted if any household member reported farming as an occupation in the Census 2011.

    e) the farming variables for the de-identified farm were then appended to every member of the ‘farm’s’ Census 2011 household so that e.g. even children’s individual records include information about farm type, livestock numbers etc.

As noted in Section 1, ‘farmers’ who chose not to claim subsidies for which they were eligible, did not meet the eligibility criteria for subsidies or lived at a second address not listed in the RPWD application will not be included in the cohort.

Farm businesses may make their RPWD application through an agent. The RPWD ‘Role’ variable (ROLE_F) contains free-text, self-reported information about the role of the person who completed the application. In order to remove individuals not believed to live in the farm household, the Role variable was used to remove e.g. ‘agents’ or ‘secretaries’ from farming households. Where an ‘agent’ or ‘secretary’ was found in the Census 2011 to be resident in a household with a self-reported ‘farmer’ they were retained in the cohort.

Figure 2: Diagram of how the ARCW Datasets were linked in the SAIL Databank


DG|CYA
Data Gweinyddol
Casgliad Ymchwil Amaethyddol
AD|ARC
Administrative Data
Agricultural Research Collection
ADR UK
Economic and Social Research Council
ARC-W Data Linkage in SAIL
FILE 1: Agricultural Data ALFs (+ Encrypted CRN and CPH)
FILE 1: Agricultural Data RALFs (+ Encrypted CRN and CPH)
2011 Census ALFs
WDSD (RALF to ALF(s))
2011 Census attribute data
WDSD e.g. Week of birth, Date of Death, Residential History
FILE 2: RPA Attribute Data + Encrypted CRN
Census Data for all Farming Household Residents
FILE 2: EUFSS Attribute Data + Encrypted CPH
AD|ARC Household level data set
Household_ID
Linkable via Household ID
AD|ARC Individual level dataset
ALF
RALF
Household_ID
Census person_ID
Linkable to other datasets in the SAIL Databank via ALF

3.3 Aggregation of agricultural data to household level

As noted above, multiple County Parish Holdings (CPHs) and farm businesses (CRNs) can be linked to the same farm household. The EU Farm Structure Survey and Rural Payments Wales variables have therefore been aggregated to the farm household level. Derived variables for a count of the number of CPHs (CPH_NUM) and CRNs (CRN_NUM) for each household are included in the dataset (see Data Dictionary for more information).

Farm Type has been recalculated at a household level using the same Robust type 10 calculations and methodology used to calculate Farm Type in the EU Farm Structure Survey.

Because, as noted above, some farm businesses are linked to more than one household, caution is required when aggregating farm business variables.

In addition to the core AD|ARC Wales Tables listed above, separate tables have been deposited in SAIL containing the complete RPWD and EUFS 2010 data for Wales. If researchers are interested in disaggregating to below the household level, e.g. to look at the values for each CRN within a household separately, it would therefore be possible to apply to link either the AD|ARC Individual-level or Household-level Table to the RPWD or EUFS tables in SAIL.

4 Key Considerations for Research using the AD|ARC Wales Database

All data has been linked and stored in either the Individual-level or the Household-level Table ready for analysis (see accompanying Data Dictionary). As noted above, additional linking may be required for more complex research questions but potential researchers should contact the core AD|ARC team who can advise on whether additional links are necessary and, if so, how to perform them.

Linking rate plus any examination of bias in linking rate to follow.

Administrative data refers to information collected for non-statistical reasons, either to enable the delivery of a public programme or service or to maintain records. Administrative data sources can be rich sources of information that can be used in quantitative analysis and evaluation without imposing an additional burden on data subjects or additional costs to data controllers. However, because they are not designed for research purposes, they often have limitations.

Although the AD|ARC project team has undertaken the steps documented above to create a Research Ready Dataset, the following limitations should be kept in mind when both undertaking analysis and reporting findings:

- The dataset is 'research ready' in that we have linked together several different datasets to enable research. The data is, however, not fully 'analysis ready'. Researchers will still be required to assess the quality of the data and consider any necessary data preparation steps in accordance with their own research needs.
- The cohort should not be seen as a census of all 'farming' households. As noted in Sections 1 and 2.1, the cohort includes all members of households in Wales that have been successfully linked to farm businesses claiming under the Single Payment Scheme in 2010. 'farmers' who chose not to claim subsidies for which they were eligible, did not meet the eligibility criteria for subsidies or lived at a second address not listed in the RPWD application will not be included in the cohort. This means that some individuals or households selected for the various AD|ARC control groups who have not stated a main occupation of farming in the 2011 Census may be involved in farming activities. This should be kept in mind when interpreting findings.
- Some farm businesses are linked to more than one household. Caution is therefore required when aggregating farm business variables.
- Farm businesses may make their RPWD application through an agent. The RPWD 'Role' variable (ROLE_F) contains free-text, self-reported information about the role of the person who completed the application. In order to remove individuals not believed to live in the farming household, the Role variable was used to remove e.g. 'agents' or 'secretaries' from farming households. Where an 'agent' or 'secretary' was found in the Census 2011 to be resident in a household with a self-reported 'farmer' they were retained in the cohort.
- Agents have been excluded (see previous point on the Role variable) unless they were also identified as being resident in 'farming' households. Therefore, commercial farms or farms conducting applications solely through an agent are unlikely to have been included in the cohort.

- There are known limitations to the construction of households using the WDSD, which is based on GP registrations. These include:

    - The absence of some or all household members due to slow reporting of address change by GP patients to their GP practice, including delays in registering with a new practice when people move house;
    - The presence of additional, ‘ghost’ household members where someone has remained registered with a GP practice at an old address, despite having moved either temporarily or permanently away;

  Due to the above – and other – limitations, the preference was to construct households using the Household ID variable (HOUSEHOLD_ID) in the Census 2011. However, for around 30% of households a farmer name of useable quality was NOT available in the combined EU Farm Structure Survey-RPWD record in order to make the link to the Census. In these cases, it was necessary to construct the household by using the WDSD to attach ‘ALFs’ to ‘RALFs’. Because of the WDSD limitations noted above, we have included in the final Table the RALF and LSOA from the WDSD as well as the dates individuals moved in and out, plus the LSOA recorded in the 2011 Census. This will allow researchers to assess the quality of the linked data in accordance with their own research needs.

The challenges and limitations documented above should be taken into consideration by researchers when preparing their application to use the Agricultural Research Collection – Wales (ARCW). The AD|ARC team welcome queries and/or potential collaborations prior to an application for use of the data.

When researchers are writing reports based on the analysis of AD|ARC project data and need to add data quality indicators or commentary, they should feel free to quote the text from this User Guide.

Because the AD|ARC datasets are complex and only just beginning to be used, early research projects could be considered beta-testers and should expect to find issues with the data, which the Project Team would be grateful if they would inform the team about so that, where possible, we can address them and where not we can provide appropriate documentation. Equally, this User Guide should be considered a ‘starter for ten’ and will be updated to include additional information or explanation where users ask the Project Team questions, so please do get in touch!

5 How is the Data Protected?

All the data held in the Agricultural Research Collection – Wales (ARCW) is de-identified. The associated data processing is administered under the framework outlined in the Digital Economy Act (2017) (DEA), which enables government to prepare administrative data for the purposes of research, and to provide de-identified versions of those data to researchers and projects accredited by the UK Statistics Authority (UKSA). The SAIL Databank is an accredited Trusted Research Environment and DCHW is an accredited Trusted Third Party under the DEA.

The SAIL Databank uses the Five Safes Framework to ensure the safety and security of its stored data. This is a set of principles adopted by a range of secure research environments to provide complete assurance for data owners.

The Five Safes are:

Safe People – trained and accredited researchers are trusted to use data appropriately

Safe Projects – data are only used for valuable, ethical research that delivers clear public benefits

Safe Settings – access to data is only possible using secure technology systems

Safe Outputs – all research outputs are checked to ensure they cannot identify data subjects

Safe Data – researchers can only use data that have been de-identified

The Agricultural Research Collection – Wales (ARCW) is stored in the SAIL Databank. For more information on how data is de-identified and protected in the Agricultural Research Collection – Wales (ARCW) please see the SAIL Databank Governance section of the SAIL website.

6 How do I access the data?

Any researcher interested in accessing the AD|ARC RRDs is advised to contact the AD|ARC Project Team to discuss their proposal.

To access the data, researchers must follow the stages under case 2 (“A research project that uses ONLY existing data in SAIL Databank”) as set out on the SAIL website.

Prior to an application to SAIL, if not already DEA accredited, researchers will be required to complete the ONS accredited researcher training.