ADR NI DATA MATCHING

1. **INTRODUCTION**

The Administrative Data Research UK (ADR) is a UK-wide partnership between universities, government departments and agencies, national statistics authorities, the third sector, funders and researchers.

The network provides a bespoke service for researchers so they can carry out social and economic research using administrative data − research which has the potential to benefit society.

ADR UK consists of four ADR Centres located in England, Scotland, Wales and Northern Ireland. Further information on the ADRN initiative can be found at <https://www.adruk.org/>

Although the ADR UK can facilitate research of singular administrative data, it is more likely that research will require linkage between various data collections from a variety of sources. This paper provides information on matching methodology used by the Trusted Third Party (TTP) in ADR NI.

1. **DATA MATCHING BACKGROUND**

NISRA has been carrying out matching exercises between administrative sources for over ten years. They have built up a wealth of knowledge and experience in this area and developed data matching techniques and methodologies based on the range of data made available to them. Details of matching that NISRA undertakes for the Northern Ireland Longitudinal Study (NILS) can be found in the [NILS Data Matching Methodology Paper](https://www.nils-rsu.co.uk/app/uploads/2019/11/Matching_Methodology_3.2.pdf) from the [NILS website](https://www.nils-rsu.co.uk/).

The objective of any data matching exercise is to join data together based on distinct identifiers for common entities between the data. Ideally the use of a common unique identifier across all systems would simplify the data linkage process. Most administrative data systems do not have common unique identifiers and therefore records have to be matched based on the available information (usually name, address and date of birth).

An exact match can be created when the details on an individual are identical in both systems. However, there are many reasons why the records for an individual on different sources may not be exactly the same. Information such as names and dates of birth may be incorrect or held inaccurately in one or both sources, and alternatives such as middle names in place of forenames or aliases in place of forenames may be present. The method of collection, purpose of the data and the amount of verification applied before the data is entered onto the systems invariably mean that the records will not be identical on all systems. The differences may be minor, such as a slightly different spelling of the surname or they could be significant such as an incorrect address, date of birth or a middle name used in place of a forename. It is therefore important to ensure that any matching methodology takes account of non-exact matched records.

1. **ADR NI MATCHING METHODOLOGY**
   1. **Requirements**

The nature of matching work to be carried out to facilitate ADR projects has a number of requirements and constraints and therefore must be:

* Automated and repeatable;
* Accurate;
* Completed in a timely manner; and
* Flexible and adaptable to varying data nuances.
  1. **Match-Keys**

To facilitate the requirements mentioned above, a ‘rule based’ or deterministic data matching approach, referred to as Match-Keys, is used. Match-keys are created by putting together pieces of information to create unique keys that can be used for automated matching.

The variation in recording demographic information across datasets can occur in a number of different forms. A single match-key alone cannot resolve all of the differences that occur between data sources, hence the need for multiple match-keys. A series of match-keys have been developed, each of which is designed to resolve particular inconsistencies between match pairs.

The highest level of matching is exact matching which links pairs of records that are identical on all matching fields. An example of a non-exact match-key is one constructed from the first two characters of an individual’s forename and surname (Bi-grams), combined with their date of birth and postcode district.

Typically, the match-keys are processed in a stepwise manner starting with the most exact match-key and working through non-exact match-keys which reduces the amount of false matches made on ‘loose’ information. 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-key.

Annex 1 shows a sample list of match-keys along with an example using fictitious information. Although the match-keys shown can be generically applied, the methodology is adapted for each project to meet the needs of differing datasets.

* 1. **Other Methodology**

The standard match-key methodology will be sufficient for most ADR NI projects. However, in some cases, depending on the characteristics of the data to be matched, other methodologies are available. These include:

* **Associative Matching** - can be used when other associated person information is available in both datasets e.g. Census Other Household Occupants, GRO Births Mother, Father, Baby. A matching routine uses basic paired information (name, DOB, gender) of each potential match subject and other household occupant. This information is then used to join both datasets to identify potential matches.
* **Ordered Longitudinal Data Matching** – used when potential match subjects have available changing demographic information over time e.g. periodic health card registration downloads, maiden name, address changes. It is known that matching between datasets from comparable time periods yields better results than that of data with a significant time difference. Iteratively carrying out matching using all available information for the same subjects over time can increase match rates significantly.
* **Address Matching** – With an increasing demand for linking address data to the LPS Pointer dataset, NISRA have developed a deterministic address matching procedure using similar match-key methodology. This process has been tested and proven to quickly yield results of a high coverage and quality.

1. **MATCHING QUALITY**

The NISRA match-key methodology has been developed using a variety of data sources available to the TTP. This mostly pertained to work on the NILS and Census Administrative Data (CAD, formerly Beyond 2011 project).

Testing of the match-key methodology was carried out by comparing results to existing *‘gold standard’* links between datasets. In general, the match-key methodology produced comparable match rates to the ‘gold standard’ links (usually 90+%) with the majority of links being made between the same records (99.5+%).

An example of the use of match-keys can be found in the online document [‘Data Matching Using Northern Ireland Administrative Data: A Worked Example[[1]](#footnote-1)’](https://www.nisra.gov.uk/publications/data-matching-using-northern-ireland-administrative-data).

With each ADR NI project, a random sample of records matched are clerically checked to assess matching accuracy. The results of this exercise are reported in the TTP matching report.

1. **SUMMARY**

* Data matching for ADR NI projects is carried out using a deterministic or ‘rule based’ approach by creating a variety of match-Keys in each dataset.
* Using identifying information, the match-key methodology iteratively pairs records, beginning with exact matches and then relaxing different match criterion to take into account a number of data ambiguities (non-exact match-keys).
* Although the match-key method shown can be generically applied, it can be adapted to take other data specific characteristics into account.
* When considering data for matching it is important to take data time comparability into account. Match rates will be higher for data from a similar time period and coverage.
* Depending on data availability, other methodologies are available to carry out further matching work to enhance or create linkages.
* The TTP team have the experience and flexibility to adapt methodology to suit the characteristics of the data.
* The Match-Key method has been tested by comparing against existing *‘gold standard’* linkages from NILS and CAD. Results are comparable and fit for purpose.
* All matching carried out for ADR NI projects will be clerically checked to assess quality of matches.
* The matching methodologies will be continually reviewed and developments applied.

**Annex 1 Match-Key Example List**

|  |  |  |  |
| --- | --- | --- | --- |
| Match-Key | Description | Inconsistencies resolved by match-key | Example |
| 1 | Forename, Surname, DOB, Gender, Postcode | None - exact agreement | donaldduck25121920mbt999aa |
| 2 | Forename Initial, Surname Initial, DOB, Gender, Postcode Sector | Name / postcode discrepancies | dd25121920mbt999 |
| 3 | Forename Bi-gram[[2]](#footnote-2), Surname Bi-Gram3, DOB, Gender, Postcode District | Name discrepancies / movers in area | dodu25121920mbt99 |
| 4 | Forename Initial, DOB, Gender, Postcode | Name discrepancies | d25121920mbt999aa |
| 5 | Surname Initial, DOB, Gender, Postcode | Name discrepancies | d25121920mbt999aa |
| 6 | Forename, Surname, Age, Gender, Postcode | Date of birth discrepancy | donaldduck95mbt999aa |
| 7 | Forename, Surname, Gender, Postcode, Age within five years of each other | Date of birth discrepancy | donaldduckmbt999aa (95+5) |
| 8 | Forename, Surname, DOB, Gender | Movers out of area | donaldduck25121920m |
| 9 | Forename, Surname, DOB, Postcode | Gender discrepancies | donaldduck25121920bt999aa |
| 10 | Surname, Forename, DOB, Gender, Postcode (matched on Matchkey 1) | Forename / surname transpositions | duckdonald25121920mbt999aa |

1. https://www.nisra.gov.uk/publications/data-matching-using-northern-ireland-administrative-data [↑](#footnote-ref-1)
2. Bi Gram is the first two characters of a string [↑](#footnote-ref-2)