ADF Companion Document A: Standard Key Variables.

Standard keys are generated by organisations carrying out ongoing data environment analysis (scanning the data environment for new data sources). You should be aware that standard keys are generic and are set up primarily for use with licence-based dissemination of official statistics and will not be relevant to every data situation. If you are using a highly controlled access environment, or at the other end of the scale open data, or if you have data that is unusual in any way, this may not be the method to use.

However, the standard keys can be useful because if your data are not safe relative to these standards then in itself that indicates that you may have a problem, even before you consider non-standard keys.

The sets of keys presented here are subsets of those generated by the Data Environment Analysis Service at the University of Manchester using the methodology reported in Elliot et al (2011). They are focused on demographics and socio-economic variables. It should be stressed that these lists are time-dependent and are very much subject to change as the data environment changes. However, they will serve as a good starting point for considering your own data situation and its key variables.

Scenario Set A: Restricted access database linkage

Scenario A1.1: Restricted access database cross match (general)

This Scenario is based upon an analysis of the information commonly available in restricted access databases.

- Home address
- Age
- Sex
- Marital status
- Number of dependent children
- Distance of journey to work
- Number of earners
- Primary economic status

- SOCmajor (Standard Occupational code)

Attacker Profile: Person with access to restricted access dataset or hacker able to obtain such access.

Scenario A1.2: Restricted access database cross match (general, extended)

This scenario is based upon an analysis of the information commonly available in restricted access databases, a slightly extended version of B1.1 with additional, less common variables. Typical variables are:

- Age
- Sex
- Marital status
- Number of dependent children
- Workplace (typically a geographical identifier)
- Distance of journey to work
- Number of earners
- Tenure
- Number of cars
- SOCmajor
- Primary economic status
- Income

Attacker Profile: Person with access to restricted access dataset or hacker able to obtain such access.

Scenario A2.1: Restricted access database cross match (health)

This represents an attack from a restricted access dataset which also contains health information. Such datasets are becoming more common. Typical core variables are:

- Home address
- Age
- Sex
- Marital status
- Employment status
- Ethnic group
- Alcohol consumption
- Smoker/non-smoker
- Long term illness
- Type of primary long term illness (possibly match against multiple variables)

Attacker profile: Individual with access to restricted access dataset.

Scenario A2.2: Restricted access database cross match (health, extended)

This represents an attack from an extended restricted access dataset which also contains health information. Such datasets are becoming more common. Typical core variables are:

- Home address
- Age
- Sex
- Marital status
- Employment status
- Ethnic group
- Alcohol consumption
- Smoker/non-smoker
- Long term illness
- Type of primary long term illness (possibly match against multiple variables)
- Number of dependent children
- Workplace (typically a geographical identifier)
- Distance of journey to work
- Number of earners
- Tenure
- Number of cars
- SOCmajor
- Primary economic status

Attacker profile: Individual with access to restricted access dataset.

Scenario A3.1: Restricted database cross match (personnel)

This scenario is based on information commonly held in personnel databases. Typically this includes considerable detail on economic characteristics such as occupation, industry, economic status, basic physical characteristics (such as age, sex and ethnic group) and some information on personal circumstances (area of residence, long term illnesses, marital status and number of children).

- Home address
- Age
- Sex
- Marital status
- Primary economic position (filter)
- Occupation
- Industry
- Hours of work

- Migration in the last year
- Ethnic group
- long term illness
- Number of children.

Attacker Profile: Person working in personnel office of large organisation.

Scenario Set B: Publicly available information based attacks

Scenario B1.1: Commercial database cross match (common)

This scenario is based upon an analysis of the information commonly available in commercial databases. Typical variables are:

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Tenure
- Primary economic status
- Social grade
- Household composition

Attacker Profile: Person or organisation with sufficient resources to purchase lifestyle database type information.

Scenario B1.2: Commercial database cross match (superset, resource cost high)

This scenario is based upon an analysis of the information available in commercial databases. This is effectively a superset of available variables which could be exploited by a well-resourced attacker who links multiple data sources together.

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Tenure
- Accommodation type
- Primary economic status

- Social grade
- Household composition
- Religion
- Number of rooms
- Income
- Transport to work
- Highest qualification
- Long term limiting illness
- Workplace

Attacker Profile: Person or organisation with sufficient resources to purchase multiple lifestyle databases.

Scenario B2: Local search

This scenario corresponds to what might be obtained through estate agent details combined with the electoral register. The variable age and ethnic group from the electoral register that could be used in a crude form are not included in this variant. Typical variables are:

- Home address
- Accommodation type
- Sex
- Lowest floor in household
- Number of rooms
- Presence of bath
- Presence of central heating

Attacker Profile: Anyone.

Scenario B3: Extended local search

This scenario corresponds to what might be obtained through estate agent details combined with the electoral register. The variables (new voter/adult) and ethnic group that could be used in a crude form from the electoral register are included in this variant. Typical variables are:

- Home address
- Accommodation type
- Sex
- Lowest floor in household
- Number of rooms
- Presence of bath
- Presence of central heating
- Ethnic group

- Age group (new voter/adult)

Attacker Profile: Anyone.

Scenario B4.1: Public information (low resources, subgroup)

This scenario imagines an intruder who is drawing on publicly available data sources focusing on a particular subgroup or groups, and who is constrained in his/her use of resources.

- Home address
- Ethnic group (crude)
- Age
- Sex
- Qualifications
- Occupation
- Workplace

Scenario B4.2: Public information (high resources, subgroup)

This scenario imagines an intruder who is drawing on publicly available data sources focusing on a particular subgroup or groups, without effective resource constraints.

- Home address
- Ethnic group (crude)
- Age
- Sex
- Qualifications
- Occupation
- Workplace
- Tenure
- Accommodation type

Scenario B4.3: Public information (high resources, opportunistic targeting attack)

This scenario imagines an intruder who is drawing on publicly available data sources, targeting a small number of individuals, who have visibility perhaps because of media coverage, without any resource constraints.

- Home address
- Ethnic group
- Age
- Sex
- Qualifications

- Occupation
- Workplace
- Tenure
- Accommodation type
- Marital status
- Country of birth
- Religion
- Nationality

Scenario B5.1: Online data sweep (low resources, opportunistic targeting attack)

This scenario envisages somebody trawling the net for available sources of information. The status of such information is questionable since much of it is deliberately self-published. For specific individuals the list of variables may be much longer than this. However, these will be commonly obtainable from online CVs and sites such as dating sites:

- Home address
- Ethnic group
- Age
- Sex
- Qualifications
- Occupation
- Workplace
- Marital status
- Dependents (y/n)
- Religion
- Income
- Language

Scenario B6.1: Worker using information about colleagues

This scenario is based upon a study of what people commonly know about people with whom they work. Typically this includes considerable detail on economic characteristics, basic physical characteristics and some very crude information about personal circumstances. Typical variables are:

- Age
- Sex
- Ethnic group
- Occupation
- Workplace
- Distance of journey to work

- Industry
- Hours
- Economic status
- Long Term illness
- Number of children

Attacker profile: Anyone working in a large organisation.

Scenario B6.2: Nosy neighbour

This scenario encompasses information that would be relatively easy to obtain by observation of one's neighbours. Obviously this does not entail either a standard match or fishing type attack. In effect one would be fishing for one's neighbours in the dataset. However if one found a match one could use information in the dataset to determine whether it is rare or not. Typical variables are:

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Number of elderly persons
- Density (persons/rooms)
- Ethnic group
- Family type
- Accommodation type
- Lowest floor in household
- Multiethnic household
- Number of residents
- Number of rooms

Scenario B7.1: Combined public and visible sources

This is essentially the combination of nosy neighbour with publicly available information scenarios. This is quite a resource intensive attack because it involves hunting for information on a small group of people in public records. It is not likely to yield the information below on all neighbours.

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children

- Number of elderly persons
- Density (persons/rooms)
- Ethnic group
- Family type
- Accommodation type
- Lowest floor in household
- Multi-ethnic household
- Number of residents
- Number of rooms
- Qualifications
- Occupation
- Workplace
- Tenure
- Country of birth
- Religion
- Nationality

Scenario B7.2: Combined public, visible and commercial sources.

This is essentially the combination of nosy neighbour with publicly available information together with a superset of commercially available data. This implies a very well-resourced attacker who is carrying out a deep information gathering exercise on a small targeted population. Note the list of variables is more extensive than might be obtained on any restricted access database.

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Number of elderly persons
- Density (persons/rooms)
- Ethnic group
- Family type
- Accommodation type
- Lowest floor in household
- Multi-ethnic household
- Number of residents
- Number of rooms
- Occupation
- Workplace

- Tenure
- Country of birth
- Religion
- Nationality
- Number of cars
- Number of dependent children
- Tenure
- Accommodation type
- Primary economic status
- Social grade
- Household composition
- Income
- Transport to work
- Highest qualification
- Long term limiting illness

Scenario Set C: Collusive attacks

Collusive attacks are ones where the data subjects collude in providing information about themselves. These do not intrinsically constitute a set against which a data controller is legally bound to protect. However, a successful collusive attack could still carry some risk, for example in terms of reputational damage.

Scenario C1.1: Demonstrative political attack: restricted set

The assumption underlying this scenario is that a political group, such as an anti-government group, acts in collusion with a data subject for the purpose of embarrassing the Government by undermining its data collection/release activities. Imagine that the data subject provides the group with copies of the information they gave to the interviewers. This scenario could happen in a census, which is a major public investment. Here the data collection process is familiar to everyone, and colluding respondents could be prepared in advance, and be guaranteed to be in the collected data (and also in the outputs with a relatively high probability). In principle, a larger number of variables could be used, but in the restricted variant, we have avoided those that are difficult to code (such as occupation), on the assumption that the political organisation will attempt to minimise divergence to prevent the demonstration backfiring. We have also avoided those that give information about other individuals apart from the colluding agent, on the assumption that the use of such variables would go against the underlying rationale for the attack.

- Home address
- Age
- Sex
- Education
- Marital status
- Primary economic status
- Ethnic group
- Religion
- Country of birth
- Migration in the last year
- Tenure
- Long term limiting illness
- Self-reported health
- Income

Attacker Profile: Person or organisation with specific desire to cause political impact on the government.

Scenario C1.2: Demonstrative political attack: extended set

- Home address
- Age
- Sex
- Marital status
- Primary economic status
- Ethnic group
- Religion
- Country of birth
- Migration in the last year
- Long term limiting illness
- Self-reported health
- Income
- Number of rooms
- Tenure
- Housing type
- Number of residents
- Number of children

Attacker Profile: Person or organisation with specific desire to cause political impact on the government.