**Developing an ONS Population Spine**

**1.0 Background**

The ONS is developing a population spine to support transformation of outputs based on integrated census, surveys and administrative records. The concept is broadly based on approaches developed in countries that have transitioned to register-based censuses where administrative population registers and personal identification numbers (PINs) are in place to uniquely identify every citizen (past and present) residing in the country.

In recent years, countries like the UK, Canada and New Zealand have started looking into the feasibility of doing so without these preconditions, by constructing population spines from the integration of many more administrative registers and sources. Consequently, it is very important to distinguish the ONS population spine, which is being designed purely for statistical purposes, from official population registers established in countries with register-based censuses. An ONS ‘statistical’ population spine will not have the necessary wider operational procedures put in place to deliver a population spine that has a comparable level of quality in terms of coverage and accuracy.

The ONS population spine will be designed to capture some of the properties underpinning central population registers, with two distinct advantages: (1) *efficiency*; once datasets have been linked to the ONS population spine, linkages across all datasets can be easily obtained using a statistical person identifier (equivalent of a PIN), (2) *privacy*; the statistical identifier enables data to be linked without the need to retain any person identifying information on datasets accessed by researchers.

1. **Functions of the ONS Population Spine**

The ONS population spine will enable the production of official statistics and support research purposes, including access to de-identified unit record data for ONS researchers.

There are a range of potential uses in official statistics that would be enabled by the population spine, including:

* the release of demographic population estimates (by age, sex, key population characteristics and sub-national geographies)
* the basis for future Admin Data Census outputs and wider population and migration statistics
* providing the population base for statistics on various topics through data sources linked to the population spine (e. g. income statistics)
* as a survey frame of people for selecting a sample for social surveys, for example increased sample inclusion of sub-populations such as ethnic groups or new migrants
* as an auxiliary source of information to support the treatment of nonresponse in surveys or the production of small area estimates

Many of these statistics are based on the *usual resident* population, however the spine needs to be comprehensive in identifying other sub-populations of interest, including previous residents that have emigrated, short term migrants, and deceased residents.

The accuracy needed for different uses of the population spine may vary. We expect that populations derived directly based on the population spine alone will not be sufficient to meet more stringent accuracy requirements for some uses of official population statistics and therefore coverage measurements and adjustments will be needed. Some statistical adjustments are often made by countries that have centralised administrative population registers and personal identification numbers.

The design of the population spine must allow for a range of research purposes covering government policy and across many topic domains. Research will include cross-sectional analysis, using information on individuals that is linked across a variety of data sources, as well as longitudinal analysis, that follows the same individuals over time. New data sources should be easily added to the population spine as the need and opportunity arises.

The role of the population spine as a central spine enabling the linking of disparate person-level data sources is critical for nearly all research. To achieve this, the population spine should aim to include all of those individuals found on key administrative data sources. Very short-term visitors such as tourists could reasonably be excluded for most research associated with population spine. ONS have been in receipt of ongoing supply of data extracts from DWP, HMRC, NHS Digital and the DFE since dates close to the 2011 Census, which is the proposed date for how far back the population spine should aim to include all residents.

The population spine would be continually updated as new people enter the population. We use the term 'ever-resident’ to denote people who have been resident at any time over the period covered by the population spine. This ‘ever-resident’ definition of the population spine target population would meet the needs of researchers as well as official statistics production.

**3.0 ONS Reference Data Management Framework**

Fundamental to the redesign the population spine is that it should be part of a wider ONS Reference Data Management Framework (RDMF) Strategy. The population spine forms one of three central registers of the RDMF alongside the Statistical Business Register and the Address Index. Other reference datasets that are likely to be key to the RDMF include a geography index and Standard Industrial Classifications (SIC/SOC).

Integration of the statistical registers will be underpinned by commonly held identifiers between the population spine, the address index and the statistical business register. For the population spine, individual’s address information will be matched to the Unique Property Reference Numbers (UPRN), which is the primary key of the address index. The HMRC PAYE schema is also intended to provide a link between employees appearing in the population spine and employers on the business register.

While the business register and address index are already sufficiently developed to support statistical outputs, the population spine needs significant redevelopment to meet wider office requirements.

**4.0 Prototype ONS Population Spine**

A prototype version of an ONS population spine for England and Wales has been established since 2015 to produce research outputs on admin data population estimates (see [Size of the population - estimates for England and Wales](https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/sizeofthepopulation)). The methodology for constructing the prototype are provided in more detail in Annex1. The prototype is currently undergoing considerable redesign as it has been built under certain constraints in the ONS Statistical Research Environment (SRE). The SRE was initially built as a tactical solution for enabling access to key datasets supporting the Beyond 2011 Programme. One of the SRE conditions requires that all person identifying data are encrypted prior to record linkage[[1]](#footnote-1). The hashing of data limits the use of advanced record linkage techniques, and the prototype has not been built with optimised design principles.

ONS has recently developed a new Data Access Platform (DAP) which enables implementation of improved record linkage techniques to support population spine construction. We are currently in the process of establishing key design principles to underpin the future population spine, which are summarised in this paper. We would like the assurance panel to review these principles and consider whether they are in agreement with what is proposed and identify additional requirements that should be built into the design of a future population spine.

**5.0 Design Principles for the Population Spine**

The population spine is one of the three central registers within the wider register-based statistical system. It needs to be designed to meet specific purposes at the person level, as well as functioning in the broader RDMF. The following design principles are proposed to underpin the ability of the population spine to perform the functions described above.

**5.1 Data Design Principles**

**Statistical purposes only, not administrative:** All uses of the population spine are ‘statistical’ in nature, whether they are involved in the production of official statistics or undertaking research, in that they produce aggregate statistics and describe patterns in the underlying data. The population spine is a statistical register because it is designed to have good statistical properties to support statistical and research purposes. It is not an official population register as described earlier and will not be available for external agency’s operational purposes that relate to identified individuals.

* The population spine will not be made available outside of ONS for any non-statistical purpose.
* Any access to identifiable data will be only for statistical purposes and will be carefully managed. For example, the statistical activities such as record linkage procedures performed internally by ONS to create the population spine may require access to identifiable information. Other statistical activities such as sample selection from the population spine may also use identifiable data to contact the sampled units as part of a survey. However, when statistical outputs using the population spine are made available externally for statistical purposes, it will not permit the identification of individuals.
* The design of the population spine is driven by statistical concepts for defined populations, units and variables.
* The population spine does not need to be completely accurate. While efforts should be made to minimise errors and explain or quantify errors that do remain, some noise can be tolerated depending on the use, and may be adjusted for through statistical models.

**Define populations:** A primary purpose of the population spine is to manage reference populations of people. The unit for the population spine is the person. 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 could change.

The population scope is broadly defined by the population living in England and Wales[[2]](#footnote-2). To be used for official population estimates, it should include the statistical concept of *usual resident* population as a key reference population. Other purposes may require short to medium-term migrants, for example those on student or work visas of less than 12 months.

Reference populations should be easily identifiable at any given point in time, including at the time of the Census of Population. To enable this, the population spine should be updated on a regular basis to include new people entering the population and flag people who have died or left the country.

As a working definition, we define the target population for the population spine as an “ever-resident” population, including those on work or student visas and other ‘medium-term’ stays.

**Enable linkages:** As one of the three central base statistical registers, the population spine should link to the two other statistical registers representing businesses and addresses. The link between the population spine person ID and address ID is critical for producing population statistics and for any analysis of sub-national geographies. Small area geographic location is needed for area-level analysis, but a specific residential address is required to form households. The unique property reference number (URPN) for residents’ addresses will provide a link to the Address Index. The current proposal for linking to the ONS statistical business register is to use the HMRC PAYE Schema which links employees (persons on the population spine) to employers on the statistical business register.

The population spine should enable accurate and consistent linkages to other person level data sources, where missing linkages are minimised. Variables used for this linking (name, date of birth, sex and address) should ideally be available for each person on the population spine. The population spine should also link to, or otherwise allow for the grouping of individuals into groups, such as families and households, and alternative groupings.

**Known identities within a de-identified system:** A person identification number (ONS statistical identifier) is needed to ensure that people on the population spine are uniquely identified over time and new residents can be added. Names, dates of birth, and source specific identifiers (e.g. NINos and passport numbers) will be held on the population spine, but access will be highly restricted. Only staff working on updating or improving the population spine itself will require higher levels of access. Researchers will not have access to identifiable information, except under special circumstances, e.g., when contacting people selected in a survey for which the survey frame is built from the population spine. Mechanisms for applying confidentiality measures should be built into the spine, so that users can easily and safely prevent disclosure with minimal manual intervention.

**Basic population attributes:** The population spine should include variables that are important to the whole system. This includes:

* variables that uniquely identify each person
* variables used to link to other person-level datasets not used in the construction of the population spine
* variables that link to the address index to know where the person lives
* variables needed for links to the statistical business register
* key demographic characteristics such as age, sex, and key characteristics used for population statistics or for stratifying samples
* variables that indicate life events or changes in residency status (birth, death and migration events)
* time references and versioning
* any record level variables needed to implement confidentiality algorithms

It is important not ‘overbuild’ the population spine and to keep maintenance efficient. Other non-essential characteristics should not be stored on the population spine. Instead more detailed characteristics on the individuals should be stored on separate topic-specific datasets (e.g., income, employment, educational attainment).

**Accuracy:** Information included on the population spine must be of a high quality. Primarily, there should be high coverage of the target population, and minimal duplication of records for the same person. A tension exists between these two objectives when several data sources must be linked to create the register. Linking in more sources may improve coverage. However, when there is no reliable unique ID in common we have to rely on deterministic methods, probabilistic linkages, or other methods to matching person records. When there is uncertainty involved in the matching methodology, adding more sources may improve coverage, but will also introduce linkage error (see annex 1). False matches (two persons thought to be the same) result in a person not being included in the population spine, while missed matches result in duplicate records for the same person on the population spine.

Attributes stored for a given individual should be as accurate as possible since they are authoritative values used by the entire statistical system. When the same variables are collected in multiple sources, a means of resolving conflicting values will need to be developed.

Existing linkages and attribute values may be amended if new information reveals conflicts or errors, though retrospective corrections need not be adopted, (e.g. previous population extractions derived from the population spine need not be updated to reflect attribute corrections).

Quality indicators need to be defined. These include indicators for the quality of the population spine itself (e.g. the coverage of specific populations, and for linkage errors), as well as indicators for the quality of attributes. Methods for measuring the achieved quality against these indicators are also needed.

Some uses of the population spine (e.g. the census or official population statistics) will want higher quality than other uses (e.g. research into the relationships between variables). With quality indicators available, choices can be made about the suitability of the intended uses, and whether further statistical methods (such as coverage adjustment) are needed.

**5.2 System Design Principles**

**Time referenced:** Versioning should be implemented, so that a record of changes/corrections made to the population spine is maintained. Everything within the population spine, including any changes, should be time referenced where possible. Both the time of change in the system and the reference period to which the change applies must be available.

**Managed access:** Security and privacy are paramount and access to identifiable information should be controlled and monitored. Provision for varying levels of access is needed to manage different functions. ONS staff working on updating or improving the population spine itself will require higher levels of access. Researchers should never have access to identifying information such as names, and external researchers may have additional requirements or restrictions placed on access. As with other aspects of the register system, access requirements should be easily manageable and understandable.

**Stable:** First releases of official statistics may be based on incomplete data and labelled as provisional. Once final statistics are released, these should be replicable from underlying data in the population spine. Stability also supports researchers, as analyses can be updated to reflect new data, while previous analysis based on older data will remain consistent.

**Timely:** To be useful in official statistics, the register must enable the timely release of data. Timeliness can also be important for research, e. g. where timely feedback is wanted for programme evaluation. Key input data must be available quickly, and the system must be built to allow for continuous updating. New records or new information on an existing record should be added to the spine in a consistent and straightforward manner, so that the details are always as up to date as the data allows. The spine should also be easily combined with other ONS systems (e.g. sample selection and collection systems, metadata systems, access control systems).

**Structurally efficient:** It is likely that the population spine will have to deal with a large number of sources for its regular updates, and also receive a large number of requests for extractions from business areas as it will hold widely used denominators (e.g. usual resident population, migrants). A structurally efficient design and process must be developed following the principles of simplicity, flexibility, performance, and robustness to errors.

**6.0 Population Spine requirements for 2021 Census**

The prototype population spine described in Annex 1 has been developed to support ONS transformation towards an Admin Data Census and wider population and migration statistics transformation. The redesigned spine will integrate records from the 2011 Census and, potentially support integrated outputs from the 2021 Census. The current proposal is to integrate data collected from the 2021 Census at the earliest opportunity into the population spine (which by this time should be well established), so that administrative data is available to support statistical processing. Below is a list of proposed uses of administrative data to support collection and statistical processing of 2021 Census data.

*Item Imputation:* Initial research indicates that linked census and administrative data (particularly PDS and CIS data) could improve the imputation of age where it is not collected on census forms.

*Placement of imputed households:* Following estimation of non-responding households, linked administrative data, including PDS, HESA, CIS and VOA may be used to determine which locations households should be imputed in down to small areas.

*Measuring accuracy of responses:* Linked administrative data could be used to check the quality of responses to some Census questions. Examples include PDS and SC as potential sources for checking basic demographics, and tenancy deposit scheme data for tenure.

*Improve coverage of Communal Establishments:* Communal establishments (CE’s) with bed spaces higher 100 can be identified in linked administrative datasets, which may improve on classifications available on the address index. Potential sources of data include, HESA, School Boarders, and PDS.

*QA of short-term migrants:* Linked exit checks data could be used to confirm the duration of stay that migrants have indicated on Census forms

*Predicting response profiles:* Pre-populating the address frame with administrative data may have benefits for identifying areas with low response rates and targeting field force follow up.

**Annex 1**

**Current ONS Spine Prototype**

The first population spine prototype was constructed in 2015 and comprised of four administrative datasets.

* NHS Patient Register (PR)
* DWP Customer Information System (CIS)
* Higher Education Statistics Agency Data (HESA)
* English and Welsh School Censuses (SC)

The primary aim of the prototype was to link across the records held on these datasets and obtain a single deduplicated list of all persons registered. The population spine needs to have as near to complete coverage of the ‘ever-registered’ population in England and Wales, with minimum duplications.

In addition to generating a single list that has broad coverage of the population, there are two other key functions of the population spine:

1. To generate and maintain an index number (the ONS\_ID) which is assigned to all persons included on the spine;
2. To serve as a reference dataset that assigns index numbers to other person level datasets that are not used in the construction of the spine.

The current prototype distinguishes between sources used for spine construction and sources that are referenced against it. Based on the existing method, it is advantageous to minimise the number of sources used to construct the spine, to avoid linkage conflicts which are difficult to resolve. The following sources have been linked to the prototype population spine to help establish rules for determining which records are likely to be in the usual resident population.

* Personal Demographic Service (PDS)
* National Benefits Database (NBD)
* Single Housing Benefit Extract (SHBE)
* Migrant Worker Scan (MWS)
* HMRC Pay As You Earn data (PAYE)

**Prototype Method**

There are no commonly held identifiers across the four datasets used in the prototype spine construction, therefore links are identified by comparing names, dates of birth and address. Spine linkage has been undertaken so far in the ONS Statistical Research Environment (SRE), which has strict controls for ensuring privacy of data, including use of a Secure Hashing Algorithm (SHA-256) to encrypt the data prior to linkage (ONS, July 2013). A number of constraints impact on the quality of the spine.

*Quality of administrative data:* Delays in updates, particularly change of address make the identification of accurate links more difficult.

*Number of administrative records:* Some of the computational demands of blocking large numbers of candidate matches in the SRE have required a trade off in methods, favouring those that are efficient in linking records quickly to support research outputs.

*Encryption of records:* Pseudonymisation of person identifiers (name, date of birth, address) make it difficult to measure the accuracy of links and tune the methods further. Very limited data has been made available for threshold setting or clerical resolution, so there is uncertainty about the quality links made in constructing the spine.

**Linkage methods**

The linkage methodology underpinning the prototype construction has been published in detail (ONS, July 2013). The majority of links identified between datasets are made through a sequence of deterministic match-keys % (figure 1 below).

**Figure 1: Match-keys used to link datasets for prototype**

|  |  |  |
| --- | --- | --- |
| **Key** | **Type** | **Unique records on PR (%)** |
| 1 | Forename, Surname, DoB, Sex, Postcode | 100.00% |
| 2 | Forename initial , Surname initial, DoB, Sex, Postcode District | 99.55% |
| 3 | Forename bi-gram, Surname bi-gram, DoB, Sex, Postcode Area | 99.44% |
| 4 | Forename initial, DoB, Sex, Postcode | 99.84% |
| 5 | Surname initial, DoB, Sex, Postcode | 99.44% |
| 6 | Forename, Surname, Age, Sex, Postcode Area | 99.46% |
| 7 | Forename, Surname, Sex, Postcode | 99.19% |
| 8 | Forename, Surname, DoB, Sex | 98.87% |
| 9 | Forename, Surname, DoB, Postcode | 99.52% |
| 10 | Surname, Forename, DoB, Sex, Postcode (matched on key 1) | 100.00% |
| 11 | Middle name, Surname, DoB, Sex, Postcode (matched on key 1) | 99.90% |

The criteria for including each of these match-keys in the linkage algorithm is determined by the level of uniqueness, as observed in the 2011 NHS Patient Register. Each match-key is included to allow for some discrepancy that may have occurred in the recording of name, date of birth or postcode, while still retaining a high level of uniqueness in the population. Once a unique link has been made on the first match-key between two datasets, the links are set-aside and the residuals compared on match-key 2, and so on. This method generates approximately 90-95% of links that are available between two datasets. There have been limited opportunities to quality assure the match-keys method. However, based on a small number of clerical exercises the precision of links following a sequential run of all eleven match-keys is estimated to be between 99.5% and 99.9% accurate.

Precision is generally high using the match-key method because it mainly targets the identification of links where there are minor discrepancies between match pairs. For administrative data, where the capture of names and dates of birth is likely to be more accurate and complete than survey collections, a high proportion of the links available can be obtained using the match-key approach. However, recent applications of the match-keys method in Census-CCS linkage, has found that match-keys can only recover a smaller proportion of links available (approximately 80%).

To recover some of the remaining links that are still available amongst the residual unmatched records it is necessary to use score-based or probabilistic matching. In previous matching for the 2011 Census (ONS, October 2012), the Fellegi-Sunter probabilistic method was used to automate the linkage of very high scoring candidates, as well as identify likely matches to be confirmed by clerical reviewers. Critical to this method was the identification of two thresholds in the match score distribution: an upper threshold, above which all candidate pairs were automatically assigned as linked, and a lower threshold, below which all candidate pairs were automatically assigned as non-links. The middle portion of distribution between the two thresholds (or ‘grey area’) was reviewed by clerical matchers to ensure that all links identified are accurate.

To set these thresholds, detailed analysis is required to observe the accuracy of links at various points in the match score distribution. This is not possible with data encrypted in the SRE, so alternative approaches have been used up until now. A logistic regression approach, where small sets of training data have been used to fit an automated model for match-decisions was originally used (ONS, 2013). This has since been replaced by supplementing the traditional Fellegi-Sunter framework with an automated approach to threshold setting, referred to as the ‘duplicate link method’ (Blakely and Salmond, 2002). Measuring the quality of probabilistic linkage in the SRE has not been possible due to the encryption of datasets which has prevented any manual assessment of linkage quality. Probabilistic matching is certainly not optimised under SRE conditions, and we anticipate that higher levels of error are prevalent than for linkage exercises involving un-encrypted data. For this reason, ONS have developed a new Data Access Platform where linkage (including probabilistic) can be undertaken without encrypting records prior to linkage. Under these conditions, probabilistic methods will be optimised, in particular for subsequent iterations of the probabilistic spine.

The probabilistic method is only applied to residual unmatched records after the sequence of match-keys have been run. Prior to operationalising the linkage algorithm in the SRE, the method was tested against a ‘gold standard’ linked dataset undertaken by the Census Quality Assurance Matching Team in 2011. Patient Register data was linked to Census and CCS records in a total of eight local authorities, and the result compared for an assessment of linkage quality. The overall linkage accuracy when combining match-keys with probabilistic linkage is estimated to be between 0.5 and 1 percent for false positives (incorrect links), and 2 to 3 percent for false negatives (missed matches).

**Prototype Construction**

Of the four sources used to construct the prototype, none of them have full coverage of the population. We use the linkage methods described above to run separate pair-wise matching exercises between datasets, as opposed to linking records simultaneously across all datasets. A detailed description is available in [Methodology of Statistical Population Dataset V2.0](https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/methodology/methodologyofstatisticalpopulationdatasetv20). Figure 2 below shows the main stages of the process.

**Figure 2: Stages in prototype construction**

**Stage 1: Stage 2: Stage 3:**

**Pair-wise matching of datasets Integrated pair-wise match outputs Collapse into single spine**

 

*Reconcile*

*conflicts*

The conflict resolution in stage 3 is necessary for handling linkage inconsistencies that can occur when matches are made across three of the datasets. These include records of students on the HESA dataset that match to both PR and CIS, and pupils on the School Census that match to both PR and CIS. Since separate pairwise matching has been undertaken at stage 1, the following types of inconsistency arise on the spine after integrating the pair-wise match outputs:

*One-to-many inconsistency:* This is where a link is established between records on two datasets that do not correspond to the same links made on the third dataset. For example, assume source A is the CIS, source B the PR and source C HESA. We might have a match output across these sources as: A1=B1, B1=C1, but A1 =C2. One option for resolving this is to break the weakest link (lets assume A1 = C2 is the weak link) and infer a consistent three-way link between A1=B1=C1. This would leave the HESA record C2 as an unmatched record on the spine. A set of rules have been established for these scenarios to determine whether a three-way link should be inferred, or whether separate links involving two persons should be captured on the spine (A1=C2 and B1=C2).

*Many-to-many inconsistency:* This is an extension of the problem above, but includes a chain of matching inconsistencies that need to be resolved when comparisons are made between records on three or more datasets. A set of rules have been established for the numerous scenarios that can occur, however such an approach would become rapidly more complex if any additional sources (for example VISA data, as is planned) was added to the spine construction.

*Incomplete closures between links:* These are relatively straightforward decisions as to whether or not to force a link in the A1=B1, B1=C1, but A1=Cnull, scenario. In this case no link for A1 has been identified on source C, so we may allow for weaker evidence of similarity between A1=C1 to confirm that a three-way link exists between these datasets.

While rules have been established for the above scenarios, they have been implemented on encrypted data without any clerical checking to assess the accuracy of decision making. Resolving the conflicts is necessary however to ensure that each record appearing on the four datasets is only assigned one ONS\_ID. This is needed for the indexing function that the spine has in supporting integration of other person level datasets.

**Maintenance of prototype**

The prototype described above has been constructed separately for each year between 2011 and 2016. The linkage is redone entirely each year, and a comparison of the links established in each year are compared with the previous year. In some cases, links that have been missed on the spine in T2 can be re-established based on what was determined in T1, if there is high confidence in the link. This is particularly important for longitudinal analysis, for example identifying when migrants appear on administrative data and the duration of their residency in UK.

The ONS\_ID is rolled over for records that appear on the spine in consecutive years, and new ONS\_ID numbers assigned to newly appearing entries on the spine. Approximately 95% of ONS\_IDs will rollover from one year to the next.

**Methodology improvements**

ONS are still in the process of designing the population spine to meet the principles laid out in main paper. Current plans are to deliver an updated version of the population spine to support population and migration statistics research by April 2019. The updated version will target specific improvements that are achievable in the short term and will be constructed in the new ONS Data Access Platform (DAP). One of the advantages of constructing the spine in DAP is that new procedures have been put in place to preserve privacy of data, negating the need to encrypt person identifying information prior to linkage. Below is a summary of improvements targeted for April.

*Additional data sources:* Analysis of the current prototype spine identified that there is undercoverage of migrant populations, particularly those that have moved to UK more recently. Home Office Exit Checks data, which includes VISA applications for non-EEA migrants will be used as an additional source to construct the population spine. In addition, the 2011 Census will be included in the spine to improve the availability of information about the migrant stock and support research into future outputs on population characteristics.

*Optimisation of probabilistic linkage:* As described above, probabilistic linkage has only been used to establish a small proportion of links on the current prototype. New tools available in DAP will now make it possible to use probabilistic linkage earlier in the matching algorithm. This will allow for replacement of some of the weaker match-keys that are more prone to false matches. The probabilistic matching algorithms will also be supported with clerical review to identify thresholds for classifying match candidates. For the most part, linkage decisions will still need to be automated due to the size of datasets being linked. However, in some instances where the quality of the matching output needs to be very high, clerical matching may be used to minimise false links and missed links.

*Revisit methods for conflict resolution:* In this paper we have summarised some of inconsistencies that occur when linking records across three or more datasets. Some early research has been explored at ONS that moves away from pair-wise linkage of data sources in favour of *n­-*wise data sources being linked simultaneously. A linkage methodology that makes decisions about whether to link multiple sources simultaneously would alleviate some of the issues described and replace some of the complex rules for resolving match output inconsistencies.

*Establish separated functions model:* As an alternative to encrypting data, a small team will have access to identifiable data to optimise linkage accuracy in the construction of the spine. This team will assign the ONS\_IDs to person level datasets available in DAP, and researchers will be able to access those datasets in accordance with permissions granted to specific projects they are working on.

**References**

ONS, July 2013, Beyond 2011: Safeguarding Data for Research: Our Policy

ONS, July 2013, Beyond 2011: Matching Anonymous Data

ONS, October 2012, Automatic Match Rates for the 2011 Census to the Census Coverage Survey

[Blakely T](https://www.ncbi.nlm.nih.gov/pubmed/?term=Blakely%20T%5BAuthor%5D&cauthor=true&cauthor_uid=12540730)1, [Salmond C](https://www.ncbi.nlm.nih.gov/pubmed/?term=Salmond%20C%5BAuthor%5D&cauthor=true&cauthor_uid=12540730).,Probabilistic record linkage and a method to calculate the positive predictive value. [Int J Epidemiol.](https://www.ncbi.nlm.nih.gov/pubmed/12540730) 2002 Dec;31(6):1246-52.

1. Person identifying data is pseudonymised prior to linkage using the Secure Hashing Algorithm (SHA-256) [↑](#footnote-ref-1)
2. The concept of a UK wide population spine is being considered by the devolved countries in UK. [↑](#footnote-ref-2)