Cabinet Office

Simplifying the transition to Individual Electoral Registration:

Preliminary findings from the 2012 pilots exploring the potential for data matching to be used to confirm existing electors.

Foreword

Individual Electoral Registration (IER) is the biggest change to our system of electoral registration for almost a century and it is essential we get it right. Not just to tackle fraud and re-establish trust in our elections, but also to ensure that as many eligible electors as possible are registered to vote. In preparing for this change we have sought evidence from a wide range of sources to inform our thinking. This includes:

- Asking the Electoral Commission to research the completeness and accuracy of the electoral register;
- Publishing a literature review of research in this area; and
- Undertaking research to explore the barriers to registration for under-registered groups.

In 2011 we also undertook pilots looking at how comparing electoral registers with other public data bases might help to improve the completeness and accuracy of the register. These pilots taught us a lot, and led my predecessor to announce proposals to simplify the transition to IER for most citizens by matching people's entries on the electoral register with data held by the Department for Work and Pensions and, where a match was found, automatically confirming them on the register, avoiding the need for them to re-apply when we change to the new system.

However we wanted to ensure these proposaIs were tested fully before we went ahead and are therefore undertaking a second set of pilots which will confirm if this system will work. I am therefore pleased to be able to publish these preliminary findings from the pilots.

The findings provide strong support for the policy of confirmation, suggesting that:

- The majority of existing electors in the pilot (around 70%) could be confirmed through data matching with DWP data.
- We can be confident in the accuracy of these matches - the vast majority of records which matched before the canvass were subsequently confirmed in the annual canvass.
- There is the potential to further increase these match rates through the use of other local or national datasets.

This provides us with reassurance that, through Confirmation, not only can we simplify the transition for a large majority of electors, but we can be confident of a good starting level of registration for the transition. Of course these are preliminary findings which will be tested further. This will include: fully testing the IT system that will support IER; looking in more detail at the potential for using local data; and better understanding the process and resource implications for Electoral Registration Officers (EROs).

I therefore look forward to seeing the results of the full evaluation of the pilot including the Electoral Commission's statutory evaluation. I would also like to extend my gratitude for the commitment and professionalism of the Local Authorities involved in these pilots who have been essential in informing our thinking.

Chloe Smith MP, Minister for Political and Constitutional Reform

Background

In 2011 the Cabinet Office ran a set of pilots exploring whether matching entries on the Electoral Register to other trusted public data sources could help to identify individuals who are not currently registered to vote but may be eligible to do so. By providing the information to enable Electoral Registration Officers (EROs) to contact these individuals and invite them to register the overall aim of the pilots was to help improve the completeness of the register, as well as improving the accuracy of the register through the identification of potentially inaccurate registrations.

During the course of these pilots, an additional potential use for data matching was identified. The 2011 pilots demonstrated that a large proportion of individuals on the electoral register (around two-thirds) could be positively matched within data held by the Department for Work and Pensions (DWP). By using data matching to 'confirm' these electors, an opportunity to automatically transfer individuals to the new IER register, without the need to provide personal identifiers, was identified. This process of 'confirmation' has the potential to simplify the transition to IER for the majority of citizens and the likely completeness of the electoral register across the transition to IER.

However, as the 2011 pilots were not designed to test data matching for the purposes of confirmation it was recommended that further testing be undertaken, across a variety of area types, to allow differences in the confirmation rates to be explored and work to assess the accuracy of the data and match rates to be undertaken.

This paper presents the preliminary findings from the evaluation of the 2012 Data Matching Pilots. The full Cabinet Office evaluation and the Electoral Commission's statutory evaluation of the pilots focussing on confirming electors will be published by March 2013.

Methodology

Participating areas

Electoral Registration Officers from across England, Scotland and Wales were invited to participate in the 2012 data matching pilots. In total, fourteen local areas volunteered to pilot data matching for the purposes of confirmation, including authorities from across England, Wales and Scotland and a range of different authority types (e.g. large urban/rural accessible). However, it is important to note that these areas were not purposively sampled and cannot be assumed to be representative of all areas.

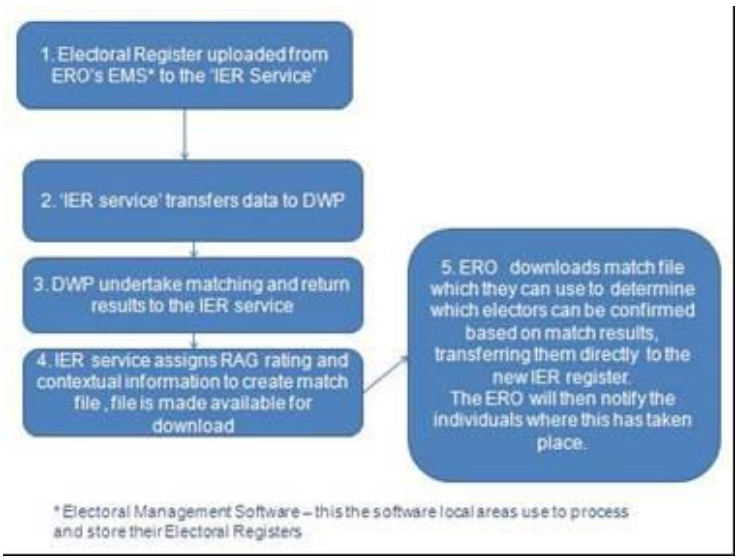
Process

The pilots sought to simulate the proposed process of confirmation, outlined in Figure 1, as far as possible.

However, as the digital solutions that will help deliver confirmation are still in the process of being developed, the pilot could not exactly replicate the process. For example, the data

was transferred via secure courier as opposed to directly through the IER digital service. This has some important implications when interpreting the results of the pilot and these are discussed more fully within the results section. Fuller testing of the digital service is planned for the full evaluation (see Box 1).

Figure 1: Outline confirmation process



Box 1: The IER digital service

The IER digital service will deliver the digital capability that enables and supports the business change from household to individual electoral registration. To date a secure environment has been built and functional components of the service have been delivered. It is now in the later stages of development and testing of the secure data transfer routes. This is on schedule to be completed in February 2013, prior to the publication of the full evaluation.

Next Steps:

The Cabinet Office will continue developing and testing components of the service using an iterative approach with GDS, Commercial EMS suppliers, EROs, the Department for Work and Pensions and other partners ready to begin roll out across local authorities from April 2013.

.

Testing accuracy

In order to improve our understanding of the accuracy of the data matching, each participating area was asked to provide two versions of their electoral register to be matched against DWP data - their pre-canvass register and their post-canvass register.

This is because the completeness of the register is known to decline during the electoral cycle (EC, 2011). Therefore, the inclusion of the pre-canvass register (which is taken at the point when it is expected that the register will be at its least complete) and the post canvass register (when it is expected that the register will be at its most complete) enables comparisons of the confirmation rate at differing points in the electoral cycle.

Furthermore, by comparing the results of the pre-canvass register to the post canvass register it is possible to assess the proportion of individuals positively matched against DWP data who were found to no longer be resident at the same address during the annual canvass. This provides an indication of the potential level of dual-inaccuracies in the matching process (i.e. where both the electoral register and DWP data may be incorrect) predominantly arising from population churn.

Finally, where pilot areas had the capacity to do so, additional data matching against local data sets was undertaken. Comparing the match results against locally held information aimed to provide further insight into both the accuracy of the data matching and whether local matching has the potential to add to the confirmation rate (by matching individuals who could not be found within the DWP data set).

Data sources & matching process

DWP Customer Information System (CIS) data

The DWP data used for the matching was a snapshot of their CIS database and is based on individuals appearing in databases kept by the Secretary of State for Work and Pensions for the purposes of social security[1]. The source CIS database is updated daily and includes a broad coverage of the population who are eligible to vote, including anyone who has been issued with a National Insurance Number (NINO). However, whilst the database has a broad coverage, it remains reliant on individuals informing DWP of changes in their circumstances (e.g. moving home). Therefore, whilst an individual may appear in the database they may not appear at their current address. This is particularly relevant for data matching for the purposes of confirmation because the limited personal identifiers available in the data mean that the matching is reliant on accurate address information[2].

A snapshot of the CIS data was extracted at a similar time as the Electoral Registers data for each pre and post canvass match run to ensure that comparable data was used. The matching itself was conducted within DWP using a matching algorithm specifically created for the pilots. The results of this

[1] CIS is also fed by a number of other Government Department feeds, primarily HMRC.

[2] The matching process works by first trying to locate the address from the electoral register within the DWP CIS data and then subsequently searching for the individual within the address. This is because the register does not currently contain additional personal identifiers that could be used for matching (e.g. date of birth), other than for attainers where date of birth is recorded.

matching were then converted into a simpler format, and assigned a basic 'Red, Amber, Green' (RAG) rating before being returned to pilot areas.

This process was carried out by the Government Digital Service (GDS) for the purposes of the pilot, using criteria developed by the Cabinet Office in conjunction with the EC and DWP[3]. This criteria was based on feedback from a number of the pilot areas, as well as learning from the 2011 pilots[4]. Full details of the matching process can be found in the match file guidance at Annex A, which was produced to assist pilot areas in interpreting the results of the data matching.

RAG categories:

- Red – no match found
- Amber – possible match found
- Green – positive match found

Locally held data sources

Where they had the capacity to do so, a number of pilot areas also used locally held data sets (for example Council Tax data or Housing Benefit data) to conduct additional data matching. This matching was conducted separately, within the individual pilot areas, and therefore the exact processes, including the matching criteria and the data sources used, will vary between areas.

Preliminary results

It is important to note that the figures presented in this report are provisional and may be subject to change as additional data becomes available and further analyses of the data are undertaken. However, this is not anticipated to impact on the overall trends reported.

Match rates

Pre-canvass match rates

Table 1 overleaf shows the proportion of entries from the pre-canvass registers of the 14 pilot areas that could be matched within the DWP CIS data. On average 71 per cent of entries could be positively matched (i.e. confirmed). The confirmation rate varies between areas, from 55 per cent in Tower Hamlets to 83 per cent in Wigan.

[3] GDS are developing the 'IER digital service', which is the IT service which supports IER and the confirmation process.
[4] In order to assist the development of the matching algorithm, match criteria (scoring algorithm) and the presentation of the match file that was returned to pilot areas, five pilot areas were used as 'Beacon' pilots. These pilot areas were sent early versions of the match files and provided feedback on the perceived quality of the matching and the usefulness of the presentation of results before the matching process. The algorithms were then 'frozen' for the purposes of the pilot, although this does not preclude further refinements of the algorithm being made prior to any wider use.

Table 1: Preliminary match rates of pre-canvass electoral registers

Pilot area	Total records matched[1]	% No match found (RED)	% Possible match (AMBER)	% Positive match (GREEN)	2011 Pilots % positive match[2]
Ceredigion	58,985	37%	2%	61%	-
Conwy	91,966	21%	3%	77%	-
Greenwich	171,905	26%	3%	71%	65%
Harrow	179,173	22%	3%	74%	-
Lothian JVB[3]	600,192	28%	3%	69%	70%
Manchester	369,996	36%	3%	61%	-
Peterborough	138,464	19%	3%	79%	-
Powys	103,072	22%	2%	76%	-
Renfrewshire JVB[4]	264,245	22%	4%	74%	-
Southwark	202,918	37%	4%	59%	-
Sunderland	218,445	17%	4%	79%	-
Tower Hamlets	171,055	40%	4%	55%	50%
Wigan	242,973	15%	2%	83%	78%
Wolverhampton	177,306	19%	3%	78%	72%
Total	**2,990,693**	**26%**	**3%**	**71%**	-

Notes: 1**.** This is the total records on the register minus any records that failed validation (i.e. were missing essential fields for matching), the number of records failing validation were very small (less than 0.5%) and would not impact on the overall proportions presented. 2. Figures are presented for those areas that participated in the pilot last year and matched their full register. They are provided for illustration only and should not be seen as directly comparable as the source data is from a different time period and a refined matching algorithm was used for the 2012 pilots. 3. Lothian Joint Valuation Board (JVB) includes the authorities East Lothian, West Lothian, Midlothian and Edinburgh, 4. Renfrewshire JVB includes Renfrewshire, East Renfrewshire and Inverclyde.

Understanding the variation in results

Findings from the 2011 pilots suggested that at least part of this variation between areas may be accounted for by differences in the population types of the areas. For example, it may be expected that areas with high population mobility (such as Tower Hamlets) will have relatively lower confirmation rates as the likelihood of either the register or the DWP CIS containing out of date or inaccurate records and therefore conflicting information may be greater. Initial results do appear to support this assumption, and further analyses of the data looking at the differences in match rates both between and within pilot areas will be conducted for the full evaluation in order to test this further.

Early analyses also suggest that the match rate varies by elector type[5]. For example, the match rate amongst postal voters was on average seven per cent higher than for non-postal voters (77% compared to 70%)[6]. The match rate for attainers[7] was also higher than for non-attainers, on average eight per cent higher (79% compared to 71%)[8]. Where possible, additional analyses of match data in order to explore the variation in match rates for different population groups will be undertaken, including for those who may be less likely to be matched. Early indications, based on feedback from pilot areas and DWP, suggest that residents of homes of multiple occupation and students may be particularly likely to fall within this group. The full evaluation will explore this further.[9]

Whilst differences in population types appear to explain at least some of the variability in match rates between areas, other factors may also play a part. For data matching to work most effectively standardisation of the data sets being matched is required. As a result, relative differences in the way that data is recorded between (and within) areas, i.e formats of the data and levels of standardisation, may impact on the effectiveness of the matching process and therefore match rates.

Whilst some standardisation of the data is automatically carried out as part of the matching process (e.g. standardising the use of 'Street', "St." and 'Str." or "Road" and "Rd") other differences may remain. In the pilot a number of differences in the way that local areas store their data were observed. As a result, GDS[10] conducted work to standardise the format of the data prior to the data being sent to DWP for matching. Primarily this work focussed on ensuring that the constituent parts of the name and/or address (e.g. address line one, address line two, county etc) were contained in the correct fields.

Should confirmation be rolled out, the planned process involves data being uploaded directly from the databases that electoral administers use to store their registers (EMS databases), through the IER digital service (the IT service which supports IER including the confirmation process), and downloaded again directly into their EMS along with the match results. The extent to which this process will be able to resolve differences in the format of the data and the subsequent impact on match rates is currently unclear. Therefore, further testing is planned which will compare the match results of data transferred though the IER

[5] These analyses exclude Lothian JVB as the data set used for the pre-canvass matching did not include elector type flags.
[6] This is the average across areas, the figures based on the total for all areas are 79% compared to69%)
[7] Attainers are 16/17 years olds who will turn 18 (and therefore become eligible to vote) within the life of the register. As attainers date of birth is held on the register to enable EROs to determine the point at which they become eligible to vote it is possible that this additional matching criteria contributed to the increased match rate, although this data was not available in all areas so cannot completely explain the trend.
[8] This is the average across areas, the figures based on the total for all areas are 81% compared to 71%.
[9] The data currently available to us does not include specific demographic data other than where it pertains to registration status (e,g, attainers). The full evaluation aims to explore differences in match rates within and across registers including at smaller geographical levels to test whether any trends can be observed.

[10] GDS are developing the 'IER digital service', which is the IT service which supports IER and the confirmation process.

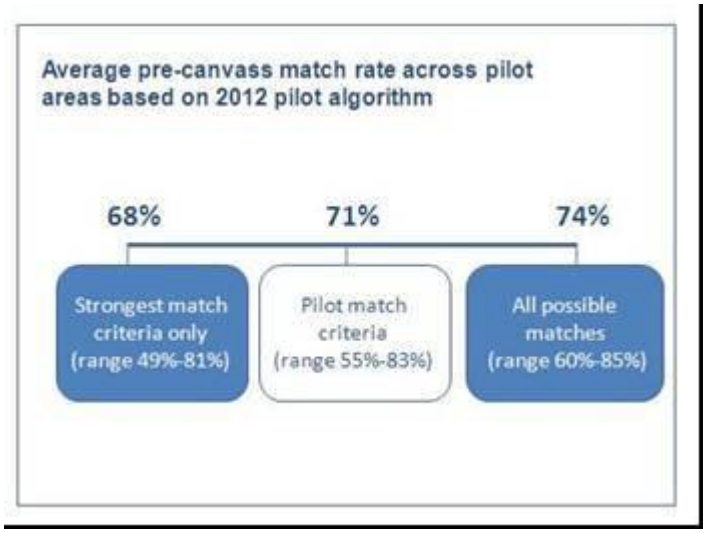
digital service end-to-end, to the results from the pilot to date.

Defining the match criteria

As described earlier the criteria for determining the thresholds for what is considered a positive match, a possible match or non match was developed for the purposes of the pilot and is detailed in Annex A. One of the aims of the pilot was to elicit feedback from pilot areas as to what is the most appropriate threshold for accepting a positive match. This will be collated through in depth qualitative interviews with pilot areas for the full evaluation. However, initial feedback does suggest that there may be some differences in views between areas as to what should be classified as a match.

To better understand the potential impact of any adjustments in the thresholds for defining what is classed as a positive match it is useful to look at the potential range of match rates produced by the pilot algorithm. Figure 2 illustrates the range of potential match rates between applying only the strongest match criteria (i.e. where the address can be found and the full first name and full surname of the individual match exactly) and the weakest match criteria (i.e. if all the 'possible matches' were assumed to be true). It illustrates that the difference between using the strongest and weakest criteria results in a 6% difference in average match rate.

Figure 2:



This means that even using the strongest match criteria on average 68 per cent of records matched in the pre-canvass register could be confirmed within DWP data. Analyses of the match results shows that across all areas, of those matched using the pilot match criteria the vast majority (96%) were matched using full first-name and surname. A further 3.5 per cent were matched using their full surname and the first three initials of their first-name. As a result further testing has been planned to explore the accuracy of the matching on this criteria as well as on the match criteria that is currently defined as 'amber' (a possible match).

These results are based on the pilot matching algorithm, however it should also be noted that DWP are continuing to carry out additional work to refine the matching algorithm using lessons learned from the pilot exercise. This has the potential to further increase the overall match rate and additional

analyses will be undertaken for the full evaluation to assess this.

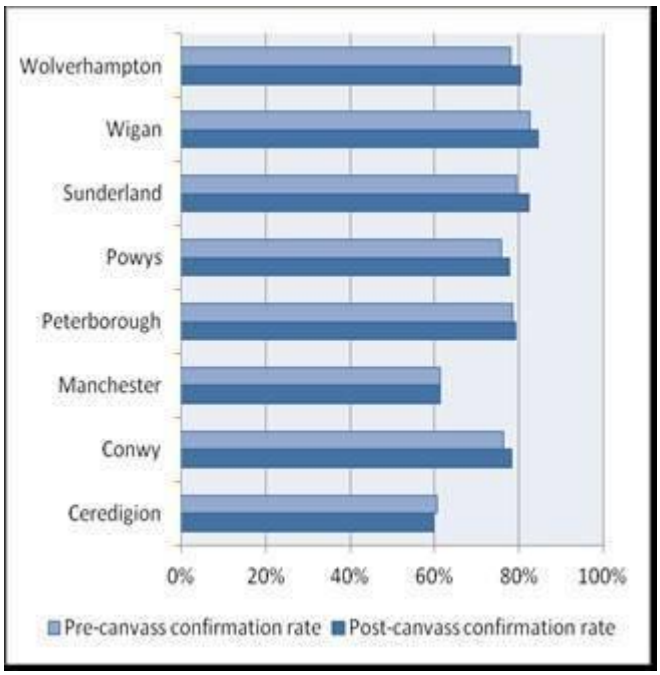
Importantly, these further analyses will seek to explore the impact of amendments to the algorithm and matching criteria on the levels of accuracy of the matching as well as the overall match rate, reflecting the importance of achieving a balance between maximising the match rate and minimising the potential degradation of the accuracy of the matching.

Post-canvass match rates

As described earlier, previous research shows that the completeness of the electoral register declines during the electoral cycle (EC, 2011)[11]. This suggests that the confirmation rate will vary according to the point in the electoral cycle in which the matching is undertaken. It is expected that the confirmation rate will be highest for the electoral register that is published following the annual canvass, where the electoral register is at its most complete. -

Usually the annual canvass runs from August to November and revised registers are published in December. This year, the first elections for Police and Crime Commissioners (PCCs) took place across England & Wales (excluding London). In these areas the annual canvass was undertaken earlier, concluding on 15 October with revised registers published on 16 October. As a result, the matching of the post-canvass registers against DWP CIS data for the eight pilot areas which held PCC elections was undertaken in October, the results of which are presented in figure 3 below.

Figure 3: Preliminary match rates of post-canvass electoral registers (PCC areas)



This shows that the difference observed between the preliminary confirmation rate of the pre-canvass and post-canvass registers was minimal. In those eight pilot areas who completed their annual canvass in October, the match rate increased in six areas (by an average of two per cent), did not change in one area and fell in one area (although by less than one per cent).

It is not possible to determine the exact reasons for this. One potential explanation is the relative currency of the data sets (i.e.

[11] EC research suggests that the completeness of the registers can be expected to decline, on average, by around one percentage point a month from the completion of the canvass.

there is a lag between an individual's details changing and DWP becoming aware of this in order to update their CIS system, meaning that DWP data is less current than the post-canvass electoral register). Further analyses of the data, including the post-canvass match results for the remaining pilot areas, will be included in the full evaluation early next year.

Accuracy indicators

Pre and Post Canvass comparisons

The proposed confirmation process enables the details of individuals on the register to be confirmed using DWP data at a particular point in time. However, as the completeness of the register is known to decline during the electoral cycle, it is expected that some individuals whose details can be found in both the electoral register and within DWP data may change their details in the period between the match being undertaken and the subsequent publication of the register (for example where an individual moves home in this period of time).

By comparing the results of the pre-canvass and post-canvass register match files it is possible to explore the proportion of individuals who were found to positively match in the pre-canvass register match but no longer matched in the post-canvass register, thereby providing an indication of the potential for this type of inaccuracy. For the eight pilot areas where the post-canvass register was available for matching with DWP data (i.e. excluding authorities in London and Scotland), an average of six per cent of individuals on the pre-canvass register were not present in the post-canvass register at the same address (range 1.5% to 8.6%).

Looking at just those individuals who were positively matched against DWP CIS data, on average, three per cent were not present in the post-canvass register at the same address (range 1.4% to 3.3%). This compares to an average of 21 per cent of individuals on the pre-canvass register who could not be matched within DWP data (range 2% to 36%).

This suggests that the risk of inaccuracies resulting from confirmation (as measured by post-canvass responses) is small, with preliminary results suggesting that around 97 per cent of individuals who were confirmed through data matching against DWP CIS in the pre-canvass register were found to still be present at the same address during the annual canvass[12].

Whilst data is not yet available for those areas without PCC elections, these areas did undertake some dip-sampling of their results during the annual canvass. These results broadly support these findings but as the

[12] This figure is based on all records, however some electors may be retained on the electoral register for a year when they have not responded to an annual canvass and we cannot be certain as to whether these records are accurate or not. We do not currently have complete data on the number of these 'carry-forward' records. However, where this data was available (for four of the eight areas) analyses showed that if you exclude carry-forward records completely, on average, 97 per cent of records that were positively matched in the pre-canvass were also positively matched in the post-canvass match file.

complete data is not yet available it is not possible to be conclusive.

Comparisons of the pre and post canvass register also enable us to see the proportion of records that cannot be matched within DWP that are subsequently confirmed in the annual canvass. As discussed previously whilst the DWP data set has a broad coverage it remains reliant on individuals informing DWP of changes in their circumstances (e.g. moving home). Therefore, whilst an individual may appear in the DWP data, they may not appear at their current address which is required for effective data matching for the purposes of confirmation[13]. As such whilst data matching against DWP can confirm a majority of electors, some electors who are accurately included in the register will not be positively matched in the data.

Overall, across the eight pilot areas for whom the post canvass match results are available, approximately 4 out of 5 electors who could not be matched in the pre-canvass register were subsequently confirmed on the register through the annual canvass.

Local Matching

As described previously, where they had the capacity to do so, a number of pilot areas also used locally held data sets (for example Council Tax data or Housing Benefit data) to conduct additional data matching. The aim of this local matching was to provide further insight into the accuracy of the data matching and to explore whether local matching has the potential to add to the confirmation rate by matching those individuals who could not be found within the DWP data set.

Early indications, based on data from four pilot areas, suggest that local data sets (primarily Council Tax records) could add in the region of ten per cent to the overall match rate[14]. Further exploration of this data is required, however, these initial results suggest that local matching could be used as a useful tool to add to the match rate, both to confirm the possible 'amber' matches and to add to the confirmation rate through confirming individuals who could not be matched (i.e. the 'red' matches).

However, it cannot be assumed that all areas currently have the technical resource available within their electoral services teams to undertake this work. This will be explored more fully with pilot areas as part of the qualitative interviews for the full evaluation.

Other national data sets

It is also possible that alternative national data sets could be used to add to the confirmation rate in the same way as local data sets. For example, indicative findings

[13] The matching process works by first trying to locate the address from the electoral register within the DWP CIS data and then subsequently searching for the individual within the address. This is because the register does not currently contain additional personal identifiers that could be used for matching (e.g. date of birth), other than for attainers where date of birth is recorded.

[14] Range 7 – 15 per cent, although further validation of these figures is required.

from the 2011 data matching pilots found that in one area, the addition of data from other national data sets similarly resulted in a rise in the overall match rate of around ten per cent[15]. However this analysis was based on one area only and therefore the results cannot be assumed to be representative and further testing would be required.

Further data matching pilots, looking at data matching for the purposes of finding new electors, are being conducted in early 2013. The results of these pilots will be used to carry out a statistical exercise to assess the potential impact of including other national data sets on the confirmation rate. It is important to note that whilst indicative findings suggests that both local and national data sources have the potential to add to the confirmation rate, it is not possible to assess the relative impact of the data sets (i.e. the extent to which the additional electors confirmed within national and local data overlap). Furthermore these findings originate from a small number of areas and therefore cannot be assumed to be representative.

Summary

The findings presented in this report are based on preliminary results only and may be subject to change as additional data becomes available and further analyses of the data are undertaken. However, it is not anticipated that the overall trends reported, which broadly support the use of data matching to confirm individuals on the electoral register, will change. Key findings include:

- Based on the average match rate achieved across the 14 pilot areas, preliminary findings suggest that around 70 per cent of records matched from the electoral register have the potential to be automatically confirmed as a result of data matching against DWP data (range 55% to 83%).

- Results from the eight pilot areas that completed their canvass in October indicate that around 97 per cent of these positive matches were also subsequently confirmed in the canvass, suggesting a high level of accuracy.

- The matching criteria (i.e. the threshold for determining what is a match) is significant for determining the match rate. For example, the difference between using the strongest and weakest criteria results in a 6% difference in the average match rate (from 68% to 74%).

- Whilst the DWP data is able to confirm a majority of electors there remains a proportion of electors who will not be able to be positively confirmed within this data despite their details being accurate on the register. Matching against other national and/or local data sets may have the potential to confirm these individuals thereby adding to the overall match rate.

[15] The full evaluation of the 2011 pilots can be accessed at http://www.cabinetoffice.gov.uk/sites/default/files/resources/FINAL-Data-Matching-Evaluation-Report-new.pdf

- Initial findings from a limited number of pilot sites supports this finding, suggesting that local matching could be used as part of the confirmation process, both as a tool for confirming possible 'amber' matches, and for adding to the match rate by confirming electors who cannot be matched within the DWP data (the 'red' matches). However, it cannot be assumed that all areas currently have the technical resource available within their electoral services teams to undertake this work.

These preliminary findings provide a valuable early indication of the potential for data matching to be used as a tool for confirmation. Additional analyses, which will be conducted for the full evaluation, will further our understanding of the potential value of confirmation and identify key lessons for implementation. This will include:

- Further analyses of post-canvass match rates, including pre and post canvass comparisons for all 14 pilot areas.

- Further analyses of the results of local matching undertaken in those pilot areas that have capacity to do so. This will include an assessment of the resources required to conduct such matching.

- Further analyses and comparisons of match rates between and within areas to enhance our understanding of the drivers of the variation in match rates.

- In depth qualitative interviews with all 14 pilot areas exploring their experience of the process of confirmation and views on the effectiveness of the pilots.

- Additional testing of the match criteria, focussing on the weaker match types and potential matches. This will be used to inform the thresholds applied for determining a match with a view to ensuring that an appropriate balance can be achieved between maximising the match rate and minimising the potential degradation of the accuracy of the matching.

- Further testing of the IT to support confirmation, including an assessment of any potential impact of using the end-to-end IT service on match rates.

Cabinet Office Data Matching Pilots 2012 – A guide to your match file

The aim of this document is to provide a guide to interpreting your match file, alongside a brief overview of the how the matching process has worked and what that means in terms of the information that we can present to you. Section 1 explains the match process and how the RAG ratings are applied, Section 2 details the individual fields in the match file, Section 3 includes some frequently asked questions and answers.

1: The Matching Process

1.1 DWP matching algorithm vs GDS scoring algorithm

Firstly, it is useful to outline the difference between the DWP matching algorithm and the GDS scoring algorithm as you may have heard us discussing the two. The initial matching of the data is undertaken by DWP using a matching algorithm created by their Information, Governance and Security Team. These match results are returned to GDS in their raw format (which is a series of coded statements about the matching). The GDS scoring algorithm transforms the data into a format that is hopefully easier to interpret and applies a RAG status to each record. This is the function that will, in the future, be carried out within the IER system (through the API). The reason that we opted for GDS rather than DWP to apply the RAG status is because it enabled us to be more flexible about altering the RAG status/presentation of data on the basis of feedback from the Beacons without requiring a formal change control process.

1.2 Address matching

Following some initial data standardisation of both the register data and DWP CIS[16] data, the DWP matching algorithm starts by trying to locate the records address in CIS This address match works differently to the identity matching described below, which you may be more familiar with. It is more like a filter (as opposed to matching on numerous criteria and then deciding the strongest match). It works as follows:

1) First it checks for a UPRN match. If a UPRN can be found then it moves on to look for an identity in that property[17].

2) If this fails it then checks for a straight match with Postcode + Address Line 1+ Address Line 2. If this passes it moves on to look for an identity.

3) If this fails it will go on instead to check for a numeric part (i.e. a house or flat number) of Address line **2**:
    a. If there is no numeric part it tries to match on Postcode + Address line 1.
    b. If there is a numeric part it matches on postcode + Address line 1 numeric + Address line 2 numeric.

    If either of these pass it is considered a match and moves on to look for an identity.

[16] CIS is the DWP Customer Information System database

[17] See additional note on UPRNs at paragraph 1.5

4) If however there is no numeric part of Address line 2 and it fails to match on the Postcode +Address line 1 criteria (3a), then it attempts to match on the Postcode + the numeric from address line 1. If this passes it is considered a match and moves on to look for an identity.

5) If all the above stages fail to return a match then it classed as a failed address match and no further matching is carried out. This means that if there is no address match, an identity match will not be attempted.

In the match file you are presented with an address match RAG which will either be 'green' (if it passed at any stage of the filter) or 'red' (if it failed at every stage of the filter). You have also been provided with the 'address match type' which tells you whether the address was matched by UPRN or address.

The key thing to note about this approach is that it means, at the lower levels of address matching (i.e. steps 3 & 4) the algorithm may be searching for an identity in more than one property, for example on a street where there are Address line 1 entries along the lines of "3 Marsden Road" and "Flat 3" which refer to different houses. This is where the combination of identity and address becomes key, as the likelihood of incorrectly getting a strong match on identity in the wrong property is thought to be relatively slim.

In addition, the number of identity matches found becomes particularly relevant. This is because, if the matching finds more than one individual on DWP CIS who appears to match at an address, this gives us a little less confidence in the accuracy of the match. In this case the GDS scoring algorithm automatically downgrades the overall RAG rating for a record to amber.

It should be noted that DWP have also done work on preparing the data before it is run through the algorithm to try and avoid some of the common pitfalls in matching as reported in last year's pilots. For example, they have prepared the data to make street endings more consistent before matching records thereby reducing mismatching where one address "Close" and the other reads "Cl". Furthermore, the filter for address focuses on numerics, which will also help with this, and for example where there is a house name in the first line of address in addition to a number.

1.3 Identity matching

The identity matching is more straight forward in terms of directly comparing each component of the name (firstname, surname etc) between the ERO record and the DWP CIS record[18]. The table below details the different combinations of identity match that are returned by DWP and the RAG status that has subsequently been attributed to them. (This is the 'Identity RAG' field on the match file). As some of the records that were supplied to DWP included DOB this is included, although in practice we will only ever have that for attainers for the purposes of confirmation.

The match file includes a separate match field for each part of the identity ('Last name match type' and so on) plus an additional field called 'Best DWP match'. The 'Best DWP match code' is simply a summary of the combination of match found (see table below), which some areas may find easier to work with.

[18] Where there were multiple ID matches found only the best match result is displayed.

Match combination	Best DWP match code	Identity RAG
DOB, LASTNAME, FIRSTNAME, MIDDLE_NAME	DOB-LN-FN-MN	Green
DOB, LASTNAME, FIRSTNAME, MIDDLE_NAME INITIAL	DOB-LN-FN-MNI	Green
DOB, LASTNAME, FIRSTNAME FIRST 3 INITIALS, MIDDLE_NAME	DOB-LN-FNI-MN	Green
DOB, LASTNAME, FIRSTNAME	DOB-FN-LN	Green
DOB, LASTNAME, FIRSTNAME FIRST 3 INITIALS	DOB-FNI	Green
DOB, LASTNAME, FIRSTNAME FIRST 3 INITIALS, MIDDLE_NAME INITIAL	DOB-LN-FNI-MNI	Green
DOB, LASTNAME, FUZZY FIRSTNAME, MIDDLE_NAME	DOB-LN-FFN-MN	Green
DOB, LASTNAME, FUZZY FIRSTNAME, MIDDLE_NAME INITIAL	DOB-LN-FFN-MNI	Green
DOB, LASTNAME, FUZZY FIRSTNAME	DOB-LN-FFN	Green
DOB, LASTNAME, MIDDLE_NAME	DOB-LN-MN	Green
DOB, LASTNAME, MIDDLE_NAME INITIAL	DOB-LN-MNI	Green
DOB, FUZZY LASTNAME, FIRSTNAME, MIDDLE_NAME	DOB-FLN-FN-MN	Green
DOB, FUZZY LASTNAME, FIRSTNAME, MIDDLE_NAME INITIAL	DOB-FLN-FN-MNI	Green
DOB, FUZZY LASTNAME, FIRSTNAME	DOB-FLN-FN	Green
DOB, FUZZY LASTNAME, FIRSTNAME FIRST 3 INITIALS	DOB-FLN-FNI	Amber
DOB, FUZZY LASTNAME, MIDDLE_NAME	DOB-FLN-MN	Amber
DOB, FUZZY LASTNAME, MIDDLE_NAME INITIAL	DOB-FLN-MNI	Amber
DOB, FUZZY LASTNAME	DOB-FLN	Amber
DOB, LASTNAME	DOB-LN	Green
LASTNAME, FIRSTNAME, MIDDLE_NAME	LN-FN-MN	Green
LASTNAME, FIRSTNAME, MIDDLE_NAME INITIAL	LN-FN-MNI	Green
LASTNAME, FIRSTNAME	LN-FN	Green
LASTNAME, FIRSTNAME FIRST 3 INITIALS	LN-FNI	Green
LASTNAME, FUZZY FIRSTNAME, MIDDLE_NAME	LN-FFN-MN	Green
LASTNAME, FUZZY FIRSTNAME, MIDDLE_NAME INITIAL	LN-FFN-MNI	Green
LASTNAME, FUZZY FIRSTNAME	LN-FFN	Amber
LASTNAME, FIRSTNAME FIRST 3 INITIALS, MIDDLE_NAME	LN-FNI-MN	Amber
LASTNAME, FIRSTNAME FIRST 3 INITIALS, MIDDLE_NAME INITIAL	LN-FNI-MNI	Amber
FUZZY LASTNAME, FIRSTNAME, MIDDLE_NAME	FLN-FN-MN	Green
FUZZY LASTNAME, FIRSTNAME, MIDDLE_NAME INITIAL	FLN-FN-MNI	Green
FUZZY LASTNAME, FIRSTNAME	FLN-FN	Amber
FUZZY LASTNAME, FIRSTNAME FIRST 3 INITIALS	FLN-FNI	Amber
FIRSTNAME AND LASTNAME REVERSED	FNLNR	Green

To note the reversal of lastname and firstname has only been used for the exact match on those 2 fields only i.e. no combination of middlename or middlename initial.

In addition, one important change to note which DWP made coming up to the code freeze, was to move away from the first name initial (1 character). Now they use the first 3 characters of the string.

A ‘fuzzy’ match in the ID match is recorded where the match has been made using ‘SOUNDEX’ which is a phonetic algorithm for indexing names by sound, as pronounced in English so that they can be matched

despite minor differences in spelling. (NOTE - The SOUNDEX algorithm is seen to be English-biased and is less useful for languages other than English)

1.4 Overall RAG rating

As described above the match file presents a separate RAG for the address and for the identity. An overall RAG is then presented, which in most cases will be the same as the identity match apart from the following exceptions:

a) Where DWP has a record of the individual being deceased ('Deceased' field) the overall rating will automatically be red.

b) Where the DWP has a record of the individual being older than 100 (Age over 100 'field') the overall rating will be red.[19]

c) Where the DWP match has returned more than one 'best match' within a property (Column X) the overall rating will be amber (as detailed earlier).

As DWP do not search for an identity where an address cannot be found within CIS, where an address match is red the overall rating will always be red.

1.5 A note on UPRN matching

As described above if the UPRN matches, no further address matches are run. Currently the UPRNs are not validated, meaning if the UPRN is assigned to the wrong address it will still pass through the matching. For the pilots we think this is acceptable, as the chances of finding someone with the same identity in the property with the incorrect UPRN is very slim. Also, the testing that some of the Beacons did suggests that the instances in which UPRNs are allocated to different address by DWP were minimal. However it should also be noted that DWP do not currently have UPRNs allocated to all addresses on their CIS database (this work is ongoing) so in some cases where you have provided a UPRN the corresponding address on DWP CIS may not have a UPRN.

We recognise the need to be a) transparent about the proportion of UPRNs assigned against both sources (ERO and CIS) for the purposes of evaluation and b) to do some further work to understand better the quality and consistency of UPRN allocation as we move forward.

[19] Originally we had anticipated this being 110 years and over, going forward we will review this to align with what is known about the oldest person alive so as not to unintentionally exclude older individuals.

2. Match file field descriptions:

Field name	Description
Ero Record Id	The unique id provided by the ERO
First Name	First name of record
Middle Names	Middle names of record
Last Name	Last name of record
Date Of Birth	Date of birth in yyyy-MM-dd
Line One	first line of Current Address
Line Two	second " " " "
Line Three	third " " " "
Line Four	fourth " " " "
Line Five	fifth " " " "
City	City of Current Address
County	County of Current Address
Postcode	Postcode of Current Address
UPRN	Unique Property Reference of Current Address
ID	Unique ID provided by the IER service
Overall RAG	The Overall RAG
Address RAG	The Address RAG
Identity RAG	The Identity RAG
Deceased	DWP have a record of indivdual being deceased
Age Over 100	DWP identified as over 100 years of age
Multiple	DWP return multiple identity matches for record
Address Match Type	Was address matched on UPRN or Address fields?
Last Name Match Type	Was lastname matched, exactly, fuzzy or not at all?
First Name Match Type	Was firstname matched, exactly, fuzzy, first 3 initials or not at all?
Middle Name Match Type	Was middlename matched, exactly, by first intitial ot not at all?
DoB Match Type	Was date of birth matched?
First/Lastname-switch	Was firstname and lastname matched reversed to achieve match
Best DWP match field	Which was the best DWP match combination (see section 1 for breakdown of codings).

3. Frequently Asked Questions:

3.1 What counts as a 'fuzzy' match in the identity match?

A 'fuzzy' match in the ID match is recorded where the match has been made using 'SOUNDEX' which is a phonetic algorithm for indexing names by sound, as pronounced in English so that they can be matched despite minor differences in spelling. (NOTE - The SOUNDEX algorithm is seen to be English-biased and is less useful for languages other than English)

3.2 How confident can we be in the matching algorithm and will this be the final version used assuming Confirmation goes ahead?

The DWP matching algorithm has been subject to internal testing and peer review within DWP providing additional confidence in the quality of the matching. Nevertheless we recognise that additional testing, and learning from the pilots, may suggest further refinements could be made.

The DWP matching algorithm was frozen on the 27th September to ensure comparability of the pre and post canvass matches for the purposes of the pilot. However, there will be an opportunity to refine the algorithm further for the purposes of any future exercises if it is deemed necessary.

3.3 Some of the records I sent don't appear to have match results, why is this?

In some cases a small number of records failed the validation to be sent to DWP (i.e. they were missing key fields such as postcode). In the future it is intended that these records would be flagged to you by your EMS system when you submit your register through the API system so that you can look into this and make any necessary amendments before submitting them, however this has not been possible at this stage.

Publication date: December 2012

This document is also available from our website at www.cabinetoffice.gov.uk
www.cabinetoffice.gov.uk