Cabinet Office

The Cabinet Office Evaluation of Data Matching Pilots, 2011

Nerissa Steel and Maria Hannan
Electoral Registration Transformation Programme

ACKNOWLEGEMENTS

The authors would like to thank all 22 pilot sites and the various data holding organisations who took part in data matching for their hard work and enthusiasm as well as the information they provided to aid the evaluation and time they took to speak to the Cabinet Office on various occasions. Similarly we would like to thank the Electoral Management System providers for supporting the individual pilot sites and aiding the evaluations. Thanks go to the Electoral Commission for their collaborative approach to the two evaluations, in particular to Gemma Rosenblatt, Phil Thompson and Davide Tiberti. Thanks are also due to IBM for providing and supporting the matching hub at the Cabinet Office.

Within the Cabinet Office we would like to thank David White and David Wilks for their work on the IT and matching involved in the pilots, to Robyn Schnuir, Marianne Ainsworth-Smith and Lester Keates for conducting the qualitative interviews and general support with the pilots and evaluation, and finally to Janet Tweedale and Mark Hughes for all their support and liaison with the individual pilot sites. Final thanks go to Dr Sarah Birch from the University of Essex and Dr Rosie Cambell from Birkbeck University for acting as independent peer reviewers of the report.

.

The authors of this report are government social researchers based in the Cabinet Office Electoral Registration Transformation Programme. The views expressed in this report are those of the authors, not necessarily those of the Cabinet Office (nor do they reflect Government policy).

All percentages reported have been rounded to whole numbers, unless otherwise stated. If you have any queries in relation to this evaluation please contact nerissa.steel@cabinet-office.gsi.gov.uk.

CONTENTS

List of tables

EXECUTIVE SUMMARY

Background

Under the current system of electoral registration an annual household canvass form is sent to each address, which is completed by one individual on behalf of everyone living at the property. From 2014 this system of registration will be replaced by one of Individual Electoral Registration (IER), with individuals registering individually and providing personal identifiers for registration.

Ensuring that the registers are as complete and accurate as possible and that levels of completeness and accuracy do not decline under IER is a key aim of the Government. Data matching, whereby records on the electoral register are matched against other sources of public data, is one tool which could assist in ensuring that the registers remain as complete and accurate as possible, both during the transition to IER in 2014/15 and on an ongoing basis.

Pilot aims

The original aim of data matching was to enable Electoral Registration Officers (EROs) to match names and addresses on their electoral register with names and addresses on existing public authority databases. Where names are found to be missing from the electoral register, EROs can offer individuals the opportunity to add their names. At the same time, if concerns are raised about a name being on the register because of fraud or error, the ERO should be able to investigate whether or not they are legitimate.

The Cabinet Office took forward 22 data matching pilot schemes in partnership with participating EROs. The pilot sites were self selecting (although the Cabinet Office selected a range of different local authority types from those who volunteered). Data sets from eight different national data holding organisations (DHOs) were matched against[1]:

- Department for Business, Innovation and Skills (BIS) - Individualised Learner Record
- Department for Education (DfE) – National Pupils Database
- Department for Work and Pensions (DWP) - Customer Information System
- Driver and Vehicle Licensing Agency (DVLA)
- Higher Education Funding Council for England (HEFCE)
- Ministry of Defence
- Royal Mail – National Change of Address file (NCOA)
- Student Loans Company (SLC) – Customer account data

Each pilot site adopted differing approaches in terms of which of these data sources they matched, the follow up actions undertaken and the groups that they sought to identify.

Research aims

This report presents the findings of the Cabinet Office evaluation of the pilot schemes. The aims of the evaluation were to examine the *process* of implementing data

[1] In addition, one pilot area matched against a local data source, the 'Citizens Account'

matching in differing local areas and to determine the *impact* of data matching on the electoral registers within the pilot areas.

The Electoral Commission (EC) has also evaluated the pilots (available at http://www.electoralcommission.org.uk/publications-and-research). This evaluation is intended to complement the EC report and the Cabinet Office has worked closely with the EC throughout the pilot.

Methodology

The over arching framework for the evaluation is based on the theory of 'realistic evaluation' which asks 'what works, for whom, under what circumstances'. The evaluation employed a mixture of both qualitative and quantitative research methods. Data was gathered from each of the pilot areas using a formal reporting form, self-evaluation reports and through qualitative interviews with each of the pilot areas. Analyses of the centrally matched data, where available, have also been undertaken. In addition to data from the pilot areas, qualitative information has been gathered from Cabinet Office staff, data holding organisations (DHOs) and software providers who have been involved in the process of setting up and running the pilots to gather learning from each of them. This was via workshops held at the Cabinet Office and short free text questionnaires.

Key findings

The process of implementing data matching

The pilot provided a valuable opportunity to test some of the processes required to enable data matching to take place effectively, and a number of key lessons were learned as a result. These are summarised below:

- A number of procedural steps are required to enable data sharing between DHOs and local areas (LAs) and to ensure data security. This process can be lengthy and it is important to ensure that sufficient time is built into the timetable of any future data matching for these activities.
- The timing of the data matching process is key. As a consequence of delays in the planned timetable for the pilot many LAs were undertaking their analysis of the data and follow up at the same time as the annual canvass. This proved highly problematic both in terms of competing resource requirements but also practical challenges in relation to completing follow up activities.
- The suggested optimum timings for a data matching exercise included either in advance of the annual canvass, in order to inform subsequent activities or directly following the canvass, when the register is at its most accurate, enabling areas to plan ongoing activities to target missing registrations. Alternative suggestions included implementing data matching on a rolling basis.
- Views on relative ease of the data transfer process were mixed. LAs may benefit from clearer guidance on the process and software requirements for data transfer, including specifications of the data required by DHOs.
- Data transfer for the pilot took place via secure email, however alternative methods of data transfer, such as a secure data hub, may offer an easier and more secure way of transferring the data.

The pilot demonstrated that the data matching process itself was complex and the volume of data returned to the LAs was reported to be much larger than anticipated owing to issues with the quality of the data, including the relative currency of data and the level of duplication within the data sets. A number of potential ways in which the data

could be improved for future exercises were identified, including:

- Applying a currency limit to the public administration data sets, to ensure that only recently active records are included within the data sets.
- Inclusion of a 'currency/activity marker' on individual records to assist LAs in identifying relative accuracy of data.
- Greater standardisation of the data sets (for example data format and match rate scores).
- Inclusion of a consistent unique identifier, namely a Unique Property Reference Number (UPRN).
- Further development of the matching algorithms to reduce the occurrence of duplication/inaccuracies within the data.
- Providing LAs with clearer guidance in relation to the data provided to them, including a description of the data and the data fields and an explanation of the matching process and match rate scores.
- Greater direct contact between the DHOs and the LAs, which may be beneficial in resolving queries and helping LAs to interpret the data sent to them.
- Providing the data in a more simple and accessible format, for example by:
  - Providing separate data files for records that are matched and those with no matches/'fuzzy' matches, or only providing detailed data on mismatches.
  - Providing data in a format that is compatible with <u>all</u> existing electoral management system (EMS) databases.
- Including nationality information in the datasets, if available, to assist LAs in identifying the records of individuals who are ineligible for inclusion on the electoral register.
- In addition, the pilot highlighted that processing and analysing the matched data requires advanced IT skills, something which was identified as a current skills gap within many electoral administration teams.

The impact of data matching

The original aims of the pilot were to test the effectiveness of the data matching in identifying missing electors (particularly amongst groups that are traditionally under-registered) and potentially fraudulent entries on the register. However the ability to robustly evaluate this has been limited by a number of factors including: the differing approaches adopted by the pilots; the level of detail included in the data sets; issues with the quality of the data; and the overlap between the annual canvass and pilot activities.

In addition to the original aims, during the course of the pilot, an alternative use for data-matching was identified, namely as a mechanism for pre-verification of individuals for the purposes of IER. Pre-verification would allow individuals whose details can be matched against trusted public datasets to be 'passported' on to the new IER register, making the process more efficient for both the individuals and EROs.

Key findings in relation to both the original aims and the potential for data matching to be used as a mechanism for pre-verification are summarised below:

- Of the data sets tested in the pilot the DWP data set had the highest match rate (the proportion of the electoral register that could be successfully matched within the national data). On average, two-thirds of the electoral register (66 per cent) could be matched within this data set.

- Combining data sets has the potential to further increase match rates. Of the data sets tested in the pilot a combination of DWP and DVLA data appears to be most effective. A detailed analysis undertaken in one LA suggests this has the potential to increase the match rate by around ten per cent although further piloting would be required to be more confident in this finding.

- Whilst the BIS, DfE and SLC data sets had much lower match rates, there is some evidence to suggest that they (BIS in particular) may be beneficial in identifying specific groups who are traditionally under-registered, namely attainers. However, further piloting with additional areas is required to test this assumption.

- Findings from the pilot suggest that Royal Mail data has the potential to identify a proportion of recent home movers, a group which has been traditionally under-registered. However, given the limited opportunity to test the data in the current pilot, further piloting of this data set is necessary to provide greater certainty of the relative benefits of its use.

- Pilot sites reported finding the MoD data to be of relatively little value owing to the lack of detail contained in the data set. A number of areas suggested that a more effective way to drive up registration rates for this group lies in effective engagement with service personnel and senior officers.

- Data matching did identify some missing electors, however the results of the pilots' follow-up work suggests that it is a less effective means of identifying and adding missing electors than the annual canvass. It is not clear the extent to which issues with the currency of the data matched and the timing of the activities impacted on these results and so further research is required to confirm this.
- The majority of pilot areas did not use the data matching as a mechanism for removing (potentially fraudulent) electors from the register. Feedback from the pilot areas suggested that this was because they did not feel confident and/or justified in using the data matching for this purpose, or that they didn't perceive fraudulent registration as an issue in their area.

Conclusions and recommendations

The pilots have provided a valuable opportunity to test the processes and effectiveness of data matching. Based on the findings, a number of recommendations for the future can be identified:

Recommendation 1: Adequate time should be allowed for the necessary legal agreements to be in place before any future data matching pilots commence, particularly where the data is viewed as being of a more sensitive nature.

Recommendation 2: Any future data matching piloting activity which requires LAs to conduct additional work needs to be considered in the context of the timing of the annual canvass and the resources available. It may be that it can offer most benefit if conducted pre-canvass in order to inform canvass activities or post canvass to identify those missing electors and check the accuracy of the register. However, some LAs may still find it beneficial to conduct matching (local matching and matching for the purposes of pre-verification in particular) during various stages of canvass activity if they desire.

Recommendation 3: Further testing and refinement of transferring data between DHOs and LAs is required to ensure the process runs smoothly.

Recommendation 4: Where possible there should be greater consistency between the national datasets and the electoral register/ EMS to ensure compatibility. In particular improved standardisation of data formats and the use of UPRNs in national datasets would improve match rates, in addition to more sophisticated algorithms.

Recommendation 5: Any future data matching should match to records which have been updated or had some activity within the previous 3, 6 or 12 months to ensure they are current and accurate. A record date should be provided and if possible the nature or reasons for the update/activity.

Recommendation 6: Any future data matching pilots should include more detailed guidance on the various datasets; what the variables mean, how they should be interpreted and used and how the matching has occurred. If possible thought should be given to involving relevant EROs and DHOs in the development of methodology at an early stage to ensure greater understanding of the data.

Recommendation 7: Further testing of some specific datasets on a larger scale, involving a consistent methodology across pilot sites is needed to see if they can effectively identify missing electors from target groups such as students, attainers and home movers.

Recommendation 8: Further testing is required on data matching for pre-verification, this should include the potential of other datasets to increase the DWP match rate, testing in a variety of area types to allow differences to be explored, and work to assess the accuracy of the data and match rates.

CHAPTER 1 - INTRODUCTION

Under the current system of electoral registration an annual household canvass form is sent to each address, which is completed by one individual on behalf of everyone living at the property. From 2014 this system of registration will be replaced by one of Individual Electoral Registration (IER), with individuals registering individually and providing personal identifiers for registration. The Government would like to ensure that the registers are as complete and accurate as possible and that levels of completeness and accuracy improve under IER. Data matching is a tool which could assist in ensuring that the registers remain as complete and accurate as possible, both during the transition to IER in 2014/15, and on an ongoing basis to supplement the canvass and possibly in the longer term, depending on its success, in place of the annual canvass.

1.1 Background

The Political Parties and Elections Act 2009 (PPE Act) put in place statutory provision for the introduction of IER in Great Britain, including a voluntary phase where Electoral Registration Officers (EROs) would invite individuals to provide identifying information but there would be no obligation for them to do so. The Coalition Agreement promised to speed up implementation of IER with the purpose of tackling electoral fraud. Therefore the Government have dropped previous plans for a voluntary phase leading up to IER and instead will legislate to bring forward implementation of compulsory IER to 2014, ahead of the next general election in May 2015. The IER White Paper was published in June 2011 for consultation. This set out that any new registrations or changes after implementation in 2014 will need to be carried out under IER and those who already appear on an electoral register will be invited to register under the new system. A carry forward arrangement has been put in place to ensure that no one who fails to register under IER will be removed from the register until after the 2015 general election. However, it will be a requirement from 2014 that anyone wishing to cast a postal or proxy vote should be registered under the IER provisions. Those who have not registered under IER will still be able to cast a vote in person but will not be able to use their absent voting methods.

The Government's response to the consultation and to the pre-legislative scrutiny of the proposals carried out by the Political and Constitutional Reform Committee was published on 9th February 2012.

Completeness and accuracy of the registers and under-registration

The Electoral Commission (EC) defines completeness and accuracy of the registers as follows:

- Completeness: 'every person who is entitled to have an entry in an electoral register is registered'.
- Accuracy: 'there are no false entries on the electoral registers'.

Measuring the completeness and accuracy of the registers is considered to be methodologically imperfect and the 'gold standard' measure is based on comparing electoral register data with the census data,

which can only be undertaken once every ten years (EC, 2010). Although, it should be recognised that even this has limitations as it is based on the quality of Census returns. The last such estimates were published in 2005 by the EC and were based on the 2000 registers; they also only covered England and Wales. The Cabinet Office therefore commissioned the EC to undertake further research into the completeness and accuracy of the registers and they contracted Ipsos-MORI to conduct a large scale national house to house survey which produced robust national estimates for the Great Britain registers as of April 2011. This study will be followed by other research into completeness and accuracy over the life time of the Programme, including a census register check which will be available in 2012/13.

The EC (2011) found that the register in April 2011 was 82% complete and 85% accurate. However, they were able to project back to December 2010 (December is the time when the register is seen to be at its most complete and accurate as this follows the annual household canvas) and found that the register was between 85 and 87% complete. This would mean that approximately 6.5 million people are missing from the electoral register. This compared to the best previously available estimate of completeness of the registers in 2000 which suggested that around 3.9 million people or 8-9 per cent of eligible voters were not registered in 2000 (EC, 2005). Completeness of the register has therefore declined over the previous ten years, making it even more important that under-registration is tackled.

Both studies found that groups such as young people (including attainers[2]), students, people who have recently moved house, people living in privately rented accommodation and/or shared households are less likely to be registered to vote. Other research (for example Fisher et al, 2011) has also highlighted lower levels of registration among the Black and Minority Ethnic (BME) population. The evidence suggests that the majority of inaccurate entries on the registers are related to people moving home and not informing the Electoral Registration Officer (EROs) (EC, 2011). It is worth noting that there is currently no requirement for people to notify the ERO when they move home which makes it more difficult for them to identify home movers. Inaccuracies linked to fraud are thought to be relatively small in number (EC, 2010), and it has been suggested that levels of inaccuracy vary in line with levels of completeness.

Data matching pilots

One possible way of helping to identify people who are currently not registered but who are eligible to be registered is via data matching (i.e. matching the electoral register against other sources of data to identify individuals and properties). This data could then be used to try and encourage these individuals to register to vote. The PPE Act therefore also provided for the creation of data matching schemes. In the autumn of 2010 the Cabinet Office produced a prospectus to invite expressions of interest from Electoral Registration Officers across Great Britain who would like to run data matching schemes. The prospectus (Annex A) set out the aims of the schemes, what would be required of those local areas which would like to be considered and the subsequent evaluation by the EC.

The Cabinet Office took forward 22 schemes in partnership with the participating EROs as part of our preparation for the introduction of IER in 2014. The pilots ran from June to November 2011. The EC has also evaluated the pilots and will report its findings by 1 March 2012 to the registration officers concerned and to the Secretary of State as well as publishing its report. The Commission's responsibility is set out in section 36 of the PPE Act; their approach to

[2] 16/17 year olds who will become eligible to vote during the life of the electoral register.

the evaluation has been developed in the context of these statutory responsibilities which state that the Commission should produce an evaluation of the pilot schemes which must assess:

- How far the schemes achieved the purpose of assisting the local registration officer to meet their objective (i.e. that people entitled to be on their register are on it; people not entitled are not on it; and that information about people who are on the register is correct);
- Whether (and if so, how much) people objected to the scheme;
- How easy the scheme was to administer; and
- Whether and how far the scheme resulted in time/cost savings.

The Cabinet Office has monitored the pilots, both to ensure they are operating successfully and that appropriate use is being made of the data produced as a result of the matching as well as to develop an early view on their likely success and to keep Ministers informed of progress. The Cabinet Office has also conducted its own evaluation of the pilots to help Ministers to decide whether data matching is something on which they might want to bring forward legislation in the future. This report sets out the Cabinet Office approach to its evaluation and the findings and recommendations from the evidence gathered[3].

1.2 Research Design

Sample

Statutory Instruments (SI) were laid which provided for data matching based on proposals made by local registration officers. The pilots were therefore self selecting (although the Cabinet Office selected a range of different local authority types from those who volunteered) and adopted differing approaches to their pilots in terms of the data sources they matched, the follow up actions undertaken and the groups that they sought to identify. This has made the evaluation and comparability of pilots and their results more complex.

Over 60 expressions of interest were received, resulting in over 40 full proposals. These proposals were reviewed at a Project Board (involving the Electoral Commission) with the decision on final selection of areas informed by several factors including geography, authority type and cost (with the aim of inclusion of a broad spread of areas) as well as the usefulness and spread of what each area were proposing to test. As illustrated in Annex B, a mix of urban and rural areas were selected, although there were relatively more large urban areas than across England/Scotland as a whole[4]. The public data sources used in the pilot were selected, by the Cabinet Office, on the basis of the extent of the coverage of the data (including amongst target groups such as students, recent movers and service personnel) and the ability to access the data within the legal timeframe for the pilot.

In order to facilitate the evaluation, pilots were required to complete a methodology questionnaire which sought both to keep the Cabinet Office updated on the progress of planning for each pilot and to help put together a meaningful framework for evaluating the pilots. The questionnaire required pilots to be clear about:

- whether the whole or part of the register was matched;
- the timing of the matching;

[3] Researchers from the Cabinet Office and the Electoral Commission have worked closely together during the course of the evaluation. Where the evaluations have approached analysis in different ways (e.g. using different data sources) these are highlighted, including the reasons for selecting the approach and impacts upon interpretation. There may be some additional minor discrepancies between the reports, for example in relation to the terminology used, however these should not impact on any of the key findings contained in this report.

[4] Large urban areas can be subject to relatively high rates of population movement and recent movers are known to be a group that is traditionally under-registered (EC, 2011) .

- the groups they expected to target;
- whether they intended on conducting any further or local matches;
- how they intend on separating the impact of the matching from the annual canvass;
- how they will be approaching the follow up work to the matching

The pilots' responses were analysed by the Cabinet Office and the EC to ensure that their approach to the matching and follow up was as robust as possible. The pilots had also been provided with a minimum standards document which had set out what was expected from each pilot by the Cabinet Office and the EC. Where their methodology was not as robust as desired they were provided with further guidance and support in order to strengthen it.

Aims and Research Questions

The overarching objective of the evaluation was:

> To examine the *process* of implementing data matching in differing local areas and to determine the *impact* of data matching on the electoral registers within the pilot areas.

This overall aim sought to assess the key issues set out above and to identify any lessons which can be learnt for policy and practice, and ultimately help inform a policy and Ministerial decision as to whether data matching should be rolled out nationally, and if so what this might look like. The evaluation seeks to answer the following research questions:

Process of data matching:

1. What processes need to be put in place before data matching can occur? E.g. legal agreements, secure email accounts in place, matching criteria agreed and written etc.
2. How long should be allowed to set up the necessary processes to provide data matching?
3. What staff, skills and infrastructure needs to be in place at a local authority and within the data holding organisation (DHO) to run a match? How does this vary by different authority types? What can be learnt from the pilots to make any roll out of data matching easier/simpler for local areas and therefore require fewer resources and less time?
4. How long did it take each authority to clean their data? This may vary depending on the particular data source and whether further matching was conducted.
5. What technical issues did the DHOs and the pilots encounter?
6. How many pilots conducted additional local matching, which local data sources were used, and what effect did this have? What was the purpose of this additional matching?
7. What approach did each pilot take to using their data and on what basis was this decision made? Did their approach vary from their initial plans, and if so why? For example, this should include any overlap with the canvass and how this was addressed.
8. What was the pilot authorities' experience of the process of data matching? And what did they feel could be improved, if anything? This information will be gathered from a primary contact in each authority (on behalf of the ERO) but will try to take into account the views of others within the authority.
9. If data matching was rolled out nationally, should this be mandatory for all local areas or should it simply be available to them should they chose to use it?

It should be noted that in terms of the process evaluation, it is not intended to be a technical evaluation of the IT infrastructure as it was not built for the purpose of a national

model, although some comment is made on the need for any specific IT support at a local level.

Impact of data matching:

10. How effective were each of the data sources in indentifying people who were not on the electoral registers who were entitled to be? How does this compare to locally held data sources where they were used?
11. How effective were each of the data sources in identifying people who should not have been on the electoral registers? How does this compare to locally held data sources where they were used?
12. How effective were each of the data sources in helping the pilot areas to assess how accurate their electoral data is? How does this compare to locally held data sources where they were used?
13. What evidence is there to suggest that data matching helps to identify and improve registration rates among traditionally under-registered groups?
14. How many people were added or deleted from the register in each pilot area as a result of data matching? How does this compare to their normal canvass activity?
15. What was the impact of data matching on the public in the pilot areas? What were their views and impressions of data matching?
16. How much did data matching cost each pilot area per elector added or removed from the register? How does this compare with the benefit that data matching delivered? A separate cost benefit analysis will be conducted by an economist for the pilots; this will have to account for the fact that there will be differences in the pilot set up costs and their scalability.
17. What is the longer term impact of data matching should it be rolled out? For example, what implications does this have for the annual canvass?

Methodology

As Ministers and policy makers are interested in both the process of data matching and the impact of data matching on the electoral register it was important that the evaluation covers both. The evaluation therefore uses both qualitative and quantitative methods to gather evidence on the implementation of data matching to evaluate the process. This has been gathered from each of the pilots using a formal reporting form (used in collaboration with the EC) and through qualitative interviews with each of the pilot areas. In addition researchers were in regular contact with each of the pilots to gather information on the process from them, for example how long it has taken them, what resources they required and how they feel that the matching has gone. The information from the informal telephone conversations has been collated and analysed to identify lessons to be learnt regarding the process.

Qualitative interviews were also conducted face to face with each of the pilot areas towards the end of their pilot to examine in more depth how the pilots have gone and gain insight into lessons that can be learnt for the future. A copy of the topic guide for the interviews can be found at Annex B. These interviews were recorded, professionally transcribed and then analysed using a thematic matrix. In addition to interviews with each of the pilots, qualitative information has been gathered from Cabinet Office staff, DHOs and software providers who have been involved in the process of setting up and running the pilots to gather learning from each of them. This was via workshops held at the Cabinet Office and short free text questionnaires sent to respondents which were analysed alongside the interviews with the pilots.

The reporting form completed by the LAs was an Excel spreadsheet which each of the pilots were asked to complete and return at key stages – after their initial analysis of their data, following their canvass and after any

follow up work they have completed. In reality the pilots sent a maximum of two versions of this form – one following initial analysis and one at the end of the pilot (some just returned a form at the end of the pilot). The form captured quantitative data on the number of records found on the different datasets which were not on their electoral register and vice versa, as well as information on the number of matches and fuzzy matches[5]. The reporting form also allowed the pilot areas to capture their thoughts on the matching process and report any successes or problems. The quantitative data was analysed using Excel.

It is important to try and isolate the counterfactual[6] and therefore every effort was made to encourage each of the pilot areas to use a control group when conducting any follow up activity during their canvass activity to make this process easier and more robust. The results of these control groups are reported where possible, but not every pilot included a control and some did not apply as robust a control as others.

Finally, at the end of their pilot each area was asked to provide their own evaluation of their pilot. This has also been analysed and included in this evaluation, this includes an assessment of how data matching has worked and the impact it has made. It also included an analysis of any feedback they have received from members of the public during the pilot, their views and reactions to data matching, as well as reasons for refusing to register when given.

In terms of an over arching framework for the evaluation, it is based on the theory of 'realistic evaluation' to help understand how local areas and ERO's will respond and implement data matching if it is rolled out. 'Realistic evaluation' asks 'what works, for whom, under what circumstances'.

Interpreting the evaluation findings

There are a number of key issues that should be taken into consideration when interpreting the findings of this evaluation:

- The 22 pilots who took forward this work were self-selecting and were encouraged to put forward their own ideas and innovations. The aim was to allow as many lessons to be learnt in relation to the policy and practice as possible, however these differences in their methodology have made the evaluation more complex.

- Owing to the timetable slippage the analysis and follow up work undertaken by the pilot sites took place at the same time as the annual canvass. As a result, where individuals have been added to the register, it has been difficult to disaggregate the impact of data matching from the standard canvass activities.

- The data sets obtained in the pilot contained limited demographic details. As a result where missing electors have been identified through data matching it has not been possible to carry out a robust assessment of who these individuals are i.e. whether they belong to those traditionally under-represented groups.

- Within the monitoring forms which pilot sites returned to the EC and the Cabinet Office detailing the results of their analysis of the data matching and the outcomes of their follow up activity, a number of inconsistencies were observed in relation to the interpretation of the variable fields and the differing match rate thresholds applied across areas.

[5] "Exact matching is very strict: either a word matches or it doesn't. An attempt to improve search recall by matching more than the exact word: fuzzy matching techniques try to reduce words to their core and then match all forms of the word". Taken from 'Expert Glossary'.
[6] What may have occurred anyway or due to other reasons/factors.

1.3 Structure of the report

The next chapter sets out the details of each pilot – the data sets they each matched, the aims of the pilots and the approach adopted for their follow up work. Chapter three examines the evidence on the process of data matching. It includes an assessment of the experiences of the process of matching for each of the pilots, the DHOs and the EMS suppliers, and how the process could be improved upon. Chapter four looks at the results of the matching process and the potential impact of data matching on the completeness and accuracy of the electoral registers, including an assessment of the usefulness of each of the different datasets. Finally, chapter five seeks to bring together the evidence gathered for the evaluation and identify the key conclusions for the study including policy recommendations.

CHAPTER 2 – OVERVIEW OF PILOTS

This chapter details the scope, aims and objectives of each of the 22 data matching pilots in England and Wales. This includes the various national data sources (and in some cases additional local data) that were matched against the electoral register in each area, the target groups (if any) that the pilots were seeking to reach and the nature and scale of the follow up work undertaken.

As detailed in the previous chapter the areas included in the pilot were those which had volunteered to participate following an invitation sent to all LAs. Areas reported a wide range of motivations for joining the pilot, both within areas as well as across areas. Most commonly these related to;

- a particular interest or perceived issue with completeness of their register relating to a specific group (e.g. students/service personnel);
- an interest in the process of data matching ;
- seeing the pilot as an opportunity to verify the accuracy/completeness of their register, test the extent of a perceived problem; and
- seeing the pilot as an opportunity to influence the policy (in some cases arising from a concern over the impact of IER) or to 'be ahead of the game'.

Following consultation with a number of EROs regarding the data sets that may be beneficial for matching against the electoral register, a number of public authorities were approached to participate in the pilot. In total eight data sets were available for matching as part of the pilot, as detailed below:

- ***Department for Work and Pensions (DWP)*** *– Customer Information System (CIS)*
  This data set is based on individuals appearing in databases kept by the Secretary of State for Work and Pensions for the purpose of functions relating to social security (i.e. claimants of working family tax credits, tax credits, child benefits, and the PAYE tax system). The source CIS database is continually updated.
- ***Department for Education (DfE) –*** *National Pupils Database (NPD)*
  This data set is based on the data included in the NPD derived from the school census, which is completed termly in January, May and October. Data on individuals in maintained schools, academies and City Technology Colleges who were at least 16 years of age but less than 19 years of age at the date the information were included.
- ***Department for Business, Innovation and Skills (BIS)***
  This dataset is based on the Individualised Learner Record (ILR) which is a collection of data about learners and their learning that is requested from learning providers in the Further Education sector and is updated at set points during the year Further information on the ILR can be accessed at http://www.theia.org.uk/ilr/

- **Higher Education Funding Council for England (HEFCE)** – This dat set is based on Higher Education Statistics Authority (HESA) individualised student record. Further information can be accessed at: http://www.hesa.ac.uk/

- ***Student Loans Company (SLC)*** This data set is based on the SLC customer accounts data, which includes records for all individuals who receive student finance i.e. loans and grants (estimated to be around 1 million students per year). It is an administrative database which is updated on a continuous basis.

- **Ministry of Defence (MoD)** Data set included details of individuals appearing in databases kept by Joint Personnel Administration (JPA), and addresses of service family accommodation managed on behalf of the Secretary of State for Defence, which appear on the database known as the ANITE housing system.

- ***Royal Mail*** – *National Change of Address file (NCOA)* Royal Mail provides a redirection service to members of the public who wish to have mail which is addressed to them forwarded to a new address. The Redirection application is verified at point of application.

  The NCOA Update and NCOA Suppress files are taken from the Royal Mail Redirections database. NCOA Update contains the names and both the new and old addresses of residential customers who have taken out a permanent redirection, and the NCOA Suppress file the moved from address only.

  NCOA Update data is made up from only those customers who have applied for a Redirection, and address details are only provided where customers have provided the relevant permission for Royal Mail to share their data, therefore it only includes data from a small section of the population, and can be used for update purpose only, not suppression. The NCOA Suppress data is made up from expired Redirections and can be used for suppression purposes only.

- ***Driver and Vehicle Licensing Agency (DVLA)*** Data set based on records of individuals in relation to whom the Secretary of State maintains driving records (as defined in section 97A of the Road Traffic Offenders Act 1988)

Each pilot site was able to select the approach that they wanted to take for the pilot, including which of these data sources they wished to match against. Table 2.1 provides an overview of the approach adopted for each pilot site. It should however be noted that in the majority of areas the approach detailed differs from the original plans set out by the pilots. This mainly arose from a delay in the exchange of data which meant that many areas were carrying out analysis and follow-up work at the same time as undertaking the annual canvass, with consequent resourcing and practicality issues. This is discussed in more detail in the following chapter.

Table 2.1 (overleaf): Overview of pilot sites

	Type of author-ity	Reg-ion	Demographics highlighted by pilot area	Data-sets used	Target groups	Matched whole/part register	Approach to follow up
Blackpool	Unitary authority	Northwest	Pockets of deprivation, profile of selected wards includes high levels of flats and houses of multiple occupation (HMOs), transient population, high benefit dependency and other social and economic problems.	• DWP • BIS • DFE	Attainers, mobile population, private renters (in areas of general under registration), non-responders (defined as those who had not returned canvass form for two years, and households identified as empty)	Sample of six particular wards with lowest response rates to canvass and the most deprived areas.	• Split sample randomly, with 50% allocated as a control group (who were canvassed in the normal way). • The other 50% were given a household visit where it was explained that their details had been identified via data matching and they were encouraged to register - they also received a follow up visit if necessary. • Those in the control group were sent a canvass form and then received a canvass visit if necessary.
Camden	London Borough Council	London	High mobility and high proportion of students, so potential for under-registration due to transient population. Also, a large private rented sector and a number of HMOs.	• DWP • BIS • DFE • SLC • HEFCE	Home movers, HMOs, students.	Whole register	• Decided that if there were up to 10,000 records which did not appear on electoral register, they would aim to follow them all up - anything above this level would be followed up using a randomising process. • Follow up would be done through personalised letters and then a visit by a canvasser to a selection of non-responders. • Follow up ran in parallel with canvass - letters were sent after the first stage and before the canvassers undertook the third stage. A sample of non- respondents were followed up by a canvasser with one visit. • Follow up focused on records which could not be imported into EMS system. Some of those were randomly selected as a control group.
Colchester	Borough Council	South East	None specifically stated.	• DWP • Royal Mail • SLC	Military personnel and their dependents, home movers and students.	Whole register (other than MoD)	• Once data was received from DWP and SLC they decided to focus follow up on 12 polling districts (approximately 20% of households and geographies of three canvassers). Areas covered were a mixture of urban, rural and suburban. • Where no household canvass was received from an address of identified individual, canvassers provided with name(s) and visited property. If not successful on the first visit, a second was made. If still no successful contact, canvasser left personally addressed letter and registration form. • Where household canvass form had been received from address of identified individual, but individual still un-registered, a personally addressed letter and registration form was sent to them. (The letter asked for reasons for not wanting to be registered if appropriate.) • For MoD data, property on-base received one household form and no further activity. Off-base property canvassed in normal way; follow up actions on-base co-ordinated with garrison.

	Type of autho-rity	Reg-ion	Demographics highlighted by pilot area	Datasets used	Target groups	Matched whole/part register	Approach to follow up
Forest Heath	District Council	East Anglia	Transient population due to the horse racing industry (high concentration of workers from the Indian Sub continent who are more difficult to contact and register, as well as Eastern European agricultural workers) and USAF personnel many of whom have British partners who will be eligible to register).	DWP	Migrant workers from Europe, spouse of USA service personnel, horse racing industry workers.	Whole register	• All individuals not included on the electoral register were followed up with a letter. • A control group of 200 was set aside.
Forest of Dean	District Council	South West	Electorate of 66,112, 1143 sixteen/seventeen year olds. 447 European citizens. 65, 504 UK Citizens, with the balance being from commonwealth countries.	• DWP • BIS • DFE • DVLA • HEFCE	Sixteen/ seventeen year olds (attainers)	Partial register - attainers only	• Non-matched records were investigated locally by using council tax records to identify any home movers. • The data was filtered to show records missing from the electoral register. • During the canvass period, the missing entries were checked against the returned household form. • Information was logged of all persons who had been included on the canvass form and would, therefore, appear on the new register. • Following the completion of the canvass, letters were sent to individuals who had been identified on the data sets, but had no entry on the elector register. A form pre-printed with name, address and date of birth was enclosed with the letter.
Glasgow	City Council	Scotland	A large student population - the wards which have the lowest return rate of canvass enquiry forms are the two where most of the students are resident in privately rented accommodation.	• DWP • SLC • DVLA	Students living in privately rented accom-odation.	Partial register - two wards	• In each of the two wards one polling district was selected for follow-up and one was selected to act as a control group (selection based on having similar make-up, namely high student population). • In the polling districts where follow up took place, enquiry forms and explanatory letters were issued addressed to the names provided from the data match. Canvassers then called at non-responding properties from both mismatches, but only at properties where no response had been received from the annual canvass. • Two individual enquiry forms plus an explanatory letter were left at each property where there was no response. At least two calls were made to each property at different times of day.

	Type of authority	Region	Demographics highlighted by pilot area	Datasets used	Target groups	Matched whole/part register	Approach to follow up
Greenwich	London Borough Council	London	Provided a lot of demographic details in the proposal regarding type of households, employment status, home ownership, and ethnicity. Borough has broad range of people including a large BME population and a fairly large student population.	• DWP • DVLA • DfE • BIS •HEFCE	Residents at addresses that have not returned their annual registration form, residents who - despite the return of annual household registration forms - are not registered, new residents, people between 17 and 30 years old, some minority ethnic groups - particularly black African nationals, military personnel and their families.	Whole register, but also have target wards for particular nationalities.	• Owing to a large volume of (probable) new identities the number of potential electors followed up was limited to approximately 12,400. • A similar sized control group was created. • Data was only included for mutually matching potential electors across the different datasets (thus improving confidence in the currency of the data and likelihood of the potential electors being actually resident). • Follow up processes consisted of writing to any individuals not registered through the canvass asking for both their household form to be completed and providing them with a personal application form and a questionnaire with a reply-paid envelope for return.
Lothian	Joint Valuation Board	Scotland	Approximately 600,000 records included on the register.	• DWP	Assessment of overall completeness & accuracy of register, with degree of focus on non-elector households & households where there are more potential electors than currently registered. Geographic analysis of results may reveal general info on, for examplw, city centre and transient populations.	Whole register	• Following analysis of the matched data two follow up groups were identified. • The first of these was referred to as "voids" - where an address shows no electors resident, but where the data match provided some names against some of these addresses. A randomly selected group was followed up on. • The second group of mismatched data included randomly selected addresses where the names shown on the DWP data were different from those held on the ERO data set. • Control groups were also created for both groups. • Action taken as follow up included a letter indicating the reasons for contact sent to relevant names and addresses. In addition a small sample of door to door canvass was carried out in the Edinburgh and Midlothian area using mismatched name data.

	Type of authority	Region	Demographics highlighted by pilot area	Datasets used	Target groups	Matched whole/part register	Approach to follow up
Manchester	City Council	Northwest	High levels of deprivation, BME, students.	• DWP	The pilot will be integrated into the ERO's new approach to the annual canvass of electors and data will be used to generate individual letters to those households which do not respond to the annual canvass. Non-responders are, therefore, the target group.	Part register - a random sample of around 10,000 (5%)	No follow up work undertaken.
Newham	London Borough Council	London	Urban area with a high turnover of population and one of the highest levels of ethnic diversity in London.	• DWP	People in privately rented accommodation, young people, home movers.	Whole register	• A sample of 1, 902 records, from a total of 20,000 was chosen for the follow-up exercise. • The records consisted of names appearing on the CIS data that could also be matched to records on the Council's CRM system. • The records on the CRM system consisted of Revenue and Benefits, Care First, Housing and Customer contact data. • Follow up letters were sent to 1, 902 persons informing them that they had been identified as being on national databases, but not currently registered on the electoral register.
Peterborough	City Council	East England	No information provided	• DWP	Aimed to target those for whom English is not their first language and HMOs.	Part register - one ward	• Intended to cross reference the match data against the latest entries from the canvass and then to follow up on any anomalies near to the closing date. • However, due to resourcing issues, the area was unable to undertake any follow-up work.
Renfrewshire	Valuation Joint Board	Scotland	No information provided.	Improve-ment service company - Citizen Account	Groups where registration levels are known to be below average, including young people and students in 18-25 age group and individuals living in areas with multiple deprivation	Whole register	• No specific follow up carried out, however, did use the canvass as an opportunity to check on properties highlighted in data matching.

	Type of authority	Region	Demographics highlighted by pilot area	Datasets used	Target groups	Matched whole/part register	Approach to follow up
Rushmoor	Shire District Council		Population 97,000, 72% of which are under 50 years old. Large military presence - garrison town in Aldershot of around 4,000 service personnel and their dependents.	• MoD	Service personnel and their dependents resident in military accommodation.	Sample of register - focus on service personnel living in military accommodation	• Data used to check existing service elector records with those held by the MoD. Where gaps were identified registrations were sought through a) sessions within the Garrison b) unit based registration events c) direct contact with individual addresses of service personnel. • In addition, arrangements were made for letters to be sent to service electors who had moved, but had valid service declaration from the Council of the MoD. • A second check of the data was completed at the end of the pilot period.
Shropshire	Unitary authority	West Midlands	Four military bases in area.	• MoD	Service personnel	Sample - service personnel and military properties (from ANITE).	• Looked at military properties and personnel. No individual follow up carried out (because of the limitations of the data).
Southwark	London Borough Council	London	Large inner London authority with a diverse population and a high population churn.	• DWP	BME, young professionals and high churn populations	Part register - three wards selected (one from each constituency and political groupings) based on inclusion of areas with significant populations of one/more of the target groups	• Data was used to support the initial canvass mail-out. This was then followed up with a mini door canvass in advance of main canvass. • During the canvass period, where no information had been forthcoming, an additional write out was undertaken.

	Type of auth-ority	Region	Demographics highlighted by pilot area	Datasets used	Target groups	Matched whole/part register	Approach to follow up
Stratford-on-Avon	District Council	West Midlands	Large proportion of retired people, also contains a munitions base with associated service personnel.	• DWP • MoD	Attainers (those who will reach 18 during the life of the register); over 70s (those aged 70 years or over on 15 October 2010); MoD personnel.	Whole register - focused on target groups	• Letters were sent out to people not on the Register and to those where there was a query, i.e. confirmation of age. • A small number of records were retained as a control group.
Sunderland	City Council	North East	Selected ward profile including 17% of the population aged over 60, 9.23% BME and 44.98% of residents classed as economically inactive.	• DWP • DfE • HEFCE • SLC • BIS	Check the accuracy of people in the benefit system to form a view as to whether this particular group is under-represented on the electoral register.	Part register - one ward of 25 chosen as sample because it is mixed in terms of political representation (Lab and Lib Dems) and has high unemployment and proportion of students. Can then compare and contrast the data.	• Where matching identified a name for a property, but no registration form had been received through the canvass, additional specifically trained canvassers went to the properties to request the information (using standard script to explain about the pilot).
Teignbridge	District Council	South West	District is heavily reliant on tourism so many rental properties are six month winter lets only.	• DWP • DVLA	Young people, transient population, benefit claimants in multiple-occupancy accommodation and elderly/vulnerable people	Partial register – wards with 5% + non response to canvass	• Due to resource constraints concentrated on DWP zero matches only. • No follow up work was undertaken.
Tower Hamlets	London Borough	London	Very diverse - almost half are BME, high levels of deprivation, significant student population and transient population.	• DWP • DfE • HEFCE • BIS	Generally under registered and houses with more than eight occupants to combat fraudulent registrations	Whole register	• Matched as many mismatches as possible against local data pre-canvass, then loaded the remainder into the canvass audit register.

	Type of authority	Region	Demographics highlighted by pilot area	Datasets used	Target groups	Matched whole/part register	Approach to follow up
Wigan	Borough Council	North West	None stated	• DWP • DVLA	Attainers, young people (18-24yrs old) and under registered	Whole register	• Concentrated only on those people appearing on external databases, but not matched against the electoral register. • Randomly selected a sample of people from records to follow-up - wrote to them explaining the reason for contacting them and inviting them to register.
Wiltshire	Shire District Council	South West	Large service voter contingent in the county	• MoD	Service Personnel	Sample - military properties only	• Compared property database of service personnel properties from the MoD to the register, so did not match individuals. • No follow up work undertaken.
Wolver-hampton	City Council	West Midlands	None stated	• DWP • DfE • HEFCE	Young people and BME groups	Whole register	• Looked at records that were missing from the electoral register data. • Selected half of the wards in the area to follow-up (by letter) with the other half acting as a control group.

CHAPTER 3 – THE PROCESS OF DATA MATCHING

3.1 - Process and logistics prior to data matching

The Cabinet Office established a number of technical options to carry out the data matching for the purpose of the pilots:

1. An ERO could send their electoral registers (or subsets of the register) direct to a DHO. The DHO would carry out the data match and then return it to the ERO. This was the case for DWP and MoD data matches. Both these organisations have data matching capability and this was a secure and efficient method.
2. EROs could receive the DHO data sets and carry out an in-house match to their records. There was some concerns on the efficiency and security of data movement using this method and it was not used during the pilots
3. A third party could carry out the data match with datasets sent to them by both DHOs and EROs. For the purposes of the pilot the Cabinet Office, with assistance from IBM, set-up a secure data matching service which used secure processes for data movement and storage.

Before any data sharing could take place between the DHO and the local areas (LAs) there were a number of standard requirements, essential activities and documents which all pilots were asked to complete or have in place due to legal requirements (i.e. Statutory Instrument or the Political Parties and Elections Act 2009). These included:

Article 4 Agreements (Information/Data Sharing Protocols)
Article 4 of the Electoral Registration Data Schemes Order 2011 required every participating ERO to make a written agreement with every DHO with which they were to match data. These agreements contained detailed information as to the respective obligations of the ERO and the DHO and set out the exact basis for the processing of data, including the requirements for the transfer, storage, destruction and security of data and the consequences of failing to meet those requirements.

The agreements were required to be signed by the ERO for each pilot site and by an appropriate official of the DHO concerned which was normally at Director level. A number of data holding organisations had also consented to the Cabinet Office matching their data on their behalf. Therefore several of the agreements were also required to be signed by Cabinet Office's Programme Director for the Electoral Registration Transformation Programme.

Completed Privacy Impact Assessment
Each of the pilot schemes were subject to their own privacy impact assessment which

sets out the details of the scheme, its effects upon individuals, security measures and their compliance with the Data Protection Act.

Data Security
There were a number of strict protocols in place to ensure the security of the data. It had been stipulated that all activities involving data conformed to all applicable legislation and to HM Government policy, including the Data Protection Act 1998 and Government Information Assurance Standards 5 and 6 (IS5 and IS6).

All staff involved in the pilot (both office staff and canvassers) who had access to information supplied by the DHO had also been required to complete information assurance training either supplied by the Cabinet Office, or an equivalent locally-arranged training which had been approved by the Cabinet Office. Every data transfer had also taken place by encrypted Secure Electronic Transfer.

Confirm compliancy with the Code of Connection
Every pilot scheme had also been required to confirm that their organisation complied with the GCSx Code of Connection (CoCo) or with the GSX CoCo in Scotland. This meant that the LAs had to confirm whether they complied with IT security standards which in turn meant that they were allowed to be connected to the Government Secure Extranet (GCSx). An exception had been made for one Local Authority. This Local Authority's level of non-compliance had been discussed with the Communications Electronics Security Group and it was agreed that it was reasonable to accept the risk and proceed.

Whilst, as detailed above, a number of procedural steps needed to be completed to enable data matching to take place, in general, interviewees from the pilot sites reported that the processes ran relatively smoothly with only minor issues experienced.

A number of respondents emphasised the value of being provided with clear guidance and support on these issues:

> *""The paperwork that I had to produce; really wasn't any trouble with that. We were given drafts and guidelines from Cabinet Office. I didn't have any issues with any of that"*

> *"I think the Cabinet Office have been really good...every time I've made a point, I've had a phone call or an email or people want to talk about it, and that's the way it should be "*

The clear exception to this related to delays in agreeing and signing the Article 4 agreements and the knock on impact on the pilot timetable, which is explored in more detail below.

Key finding: A number of procedural steps are required to enable data sharing between DHOs and LAs and to ensure data security. This process can be lengthy and it is important to ensure that sufficient time is built into the timetable of any future data matching exercises to complete this.

Pilot timing

It took the Cabinet Office longer than had originally been envisaged not only to get legal agreement from some of the data holding organisations on the Article 4 Agreements, but also the required signature from the ERO and the respective Director in the DHOs. This was exacerbated by the fact that many of the agreements were ready for signature in early summer, a period when many people were either about to go on holiday or already away on holiday. A key

consequence of this delay was that for many LAs the data matching took place at the same time as their annual canvass.

These delays impacted on some of the data sources more than others, for example the data sharing agreements with Royal Mail and MoD took the longest amount of time to agree as Royal Mail data is subject to commercial sensitivities and MoD data is very sensitive and security is of primary concern. The impact of delays in the legal agreements for these datasets was significant as it led to the data being used in a limited way, if at all due to time constraints in the period remaining for the pilots to take place.

Hitting the canvass period was consistently raised by local areas as being highly problematic both in terms of the availability of resource at this period but also practical challenges in terms of following up cases and the potential for duplication of efforts (e.g. information requested in the follow up for the pilot may also be being requested through the canvass with the risk of duplicating information requests):

> *“"running it side by side with the canvass was really difficult."*
>
> *"the closer they come together, then the harder it is to, you know, to want to send out a form when you know, next month or the month after, that you're going to be sending another form anyway to the house."*

Suggestions for the most appropriate time for a data matching exercise to take place were mostly for the period just after the register had been published (Dec/Jan/Feb). Some expressed a preference for it to take place ahead of the canvas (June/July) in order to inform canvass activities:

> *" If you want to say these are the people you need to go out and do a registration drive on, you do it on the first of January, having had the annual canvass or process thereof similar to say these are all the people you've got, these are the people you're still missing, these are the people we think from national data sets you might be missing…"*
>
> *"So, had we done it in July when we originally thought we could do it, it would have been ideal, because we would have had that and then we would have been running the canvass, post receiving that. That would have all worked quite nicely. That would have been our plan"*

Finally, the idea of implementing data matching on a rolling basis was also mooted:

> *"if this data matching comes in, I would much rather see it done on an incremental basis, change-only basis."*

Key findings: The timing of the data matching process is key, as a consequence of delays in the planned timetable for the pilot many LAs were undertaking their analysis and follow up work for the pilot at the same time as the annual canvass. This proved highly problematic both in terms of competing

The suggested optimum timings for a data matching exercise included either in advance of the annual canvass in order to inform subsequent activities or directly following the canvass, when the register is at its most accurate, enabling areas to plan activities to target missing registrations. Alternative suggestions included implementing data matching on a rolling basis.

The following sections of this chapter explore some of the issues related to transferring and processing/interpreting the data. The findings presented are generic, although where comments are specific to a single data set this is highlighted. However, it should not be assumed that all findings are relevant to all data sets.

3.2: Securely transferring the data

Having agreed the processes for sharing and matching data, the views of LAs on the relative ease of sending and receiving the data were mixed, with some areas finding the process relatively easy and others experiencing difficulties with the data transfer[7]. Where problems arose these tended to relate to data files getting caught in firewalls either as a result of their size/format or because of the requirement for files to be password protected. Some areas experienced issues using secure emails and suggested other methods of exchanging data, such as a centrally held secure hub to which data could be uploaded. In addition, some LAs had found that the software that was used by the DHOs and the LAs was not always compatible (for example differing versions of programmes).

In the majority of cases workarounds for the problems encountered were found and, as a result of this learning, should be easier to anticipate and prevent/resolve in any future exercises. Going forward providing clearer guidance on the IT requirements alongside greater standardisation of the data (discussed in more detail later in the chapter) has the potential to significantly improve this process.

It should also be noted that DWP recorded two data security incidents with regards to the transfer of data, one arising from human error and another IT related.[8] In both instances the incidents were resolved promptly, with no adverse consequences. The issues arose as a result of emailing the data to LAs and may have been avoided by using an alternative method of transferring data.

Overall, the pilots highlighted a number of lessons about the secure movement of data and how this is likely to impact on the business and technical design of future data matching exercises. In particular it will be important to maintain data security, but to avoid data matching and movement becoming an unwelcome and time consuming part of the process.

Key findings:

Views on relative ease of the data transfer process were mixed. Local areas may benefit from clearer guidance on the process and software requirements for data transfer, including specifications of the data required by DHOs.

Data security and protection is essential and pan government standards are key features required of any system. Data transfer for the pilot took place via secure email, however alternative methods of data transfer, such as a secure data hub, may offer an easier way of transferring the data whilst maintaining the data security standards required. This will also ensure consistent standards across all EROs.

[7] It should be noted that the Royal Mail and HEFCE data sets were not returned to the LA (see chapter three for further discussion) and the pilot areas matching against the MoD data did not report any difficulties with sending or receiving data, primarily owing to the relatively small size of the data set.

[8] In the first instance DWP erroneously sent one LA the data for a different LA which was immediately returned and deleted from the LA system. In the second instance the email and relevant attachments containing the data became caught in a LAs firewall which DWP were required to record as a data incident as the data was in effect missing for a short period of time.

3.3 Processing and analysing the data

Once the data had been received by the LAs, the process of processing and analysing the data highlighted a number of additional challenges and potential areas for improvement for future exercises.

In the first instance the sheer volume of data returned to LAs following the DHOs matching proved problematic. Whilst most of the pilot areas did not have any set expectations about what the data they were likely to receive would look like, many reported finding the volume of data sent back to them to be much larger than they had anticipated and, in some cases, overwhelming:

> *"It's quite overwhelming and quite scary actually when you kind of looked at their total numbers at the bottom of every spreadsheet of how many records... data there was there. "*

> *"The biggest shock was the size of, the volume of data that we got, particularly from DWP, and in the consolidated, the DVLA records. Shockingly large."*

The volume of the data appears to have be driven by three key features of the data. Firstly the currency of the data, secondly the lack of standardisation within the data sets, and thirdly levels of duplication within the datasets. These are discussed in more detail below.

Currency

Due to the nature of public data sets some records will be held on the systems regardless of the length of time since the last contact between the individual and the DHO. The longer the period since the last contact with an individual the less accurate the information held is likely to be, as individuals are more likely to have moved addresses etc over time. The consequence of this was that many records were included which were known to be out of date and included individuals who had since moved out of the area.

Early pre-piloting of the DWP data set in two areas had identified this as an issue and as a result this data set was limited to only include records where there had been some update of the record within the last two years. However, other data sets did not have a currency limit and based on their experiences of the pilot, local areas have suggested that a shorter currency limit may be required on the DWP data. It has been suggested that only data which has been updated within the last 12 months at least and preferably in the last 3 or 6 months should be included. In addition, it has been suggested that each individual record should include a 'currency/activity marker', which would display the date of when the record was last updated/active, enabling clearer comparison between data and giving users greater confidence in the accuracy of the data[9]. Key findings:

Applying a currency limit to the public administration data sets, to ensure that only recently active records are included within the data sets, has the potential to significantly improve the data matching process.

9 It had not been possible in this round of pilots to provide a record date as the statutory instrument (SI) had not actually specified this. Instead, the pilot schemes were only allowed to set a filter allowing updated records for a certain period to be returned. It will be important to ensure this is included in any future SIs to enable this information to be included.

The inclusion of a 'currency/activity marker' on individual records would greatly assist LAs in identifying relative accuracy of data.

Standardisation of data

Another key issue experienced with matching across data sets resulted from a lack of standardisation within the data sets, which led to a number of records being incorrectly identified as unmatched owing to differences in name/address conventions. For example, in some data sets a full middle name was used whilst in others only the initial was included[10].

Whilst electoral registration records are primarily address-based other public data sets are primarily individual-based, which further complicated the matching process. Almost all areas suggested that the addition of Unique Property Reference Numbers (UPRNs) to the data would have significantly improved the process in respect of this. UPRNs are standardised unique identifiers for each land and property unit and are heavily used by EROs to conduct their current activities.

The lack of standardisation of data across data sets was also reported to be an issue for LAs as they found that different data sources e.g DWP, DfE were presented in different ways making it more difficult to match across the data sets. This was further complicated by the different match processes that had been used for each of the data sets which made it more difficult to compare the match rates across data sets as they were based on different scales[11]. Standardising the format of the data was a key way in which LAs felt that the data could be improved. Some areas suggested that having a central hub into which all data could be uploaded would be beneficial which, as highlighted earlier, would have the additional benefit of facilitating the data transfer process.

Key findings:

Greater standardisation of the data sets (for example data format and match rate scores) would assist LAs in processing and analysing the matched data.

Inclusion of a consistent unique identifier, namely a UPRN, has the potential to significantly improve the data matching process.

Duplication

Linked to the above point, the high number of duplicate records within the datasets was also commonly highlighted as an issue with the data. This duplication appears to have occurred for two key reasons. Firstly, where data sets were based on multiple sources an entry could appear more than once owing to the differences in name/address conventions between the data sets. Secondly, within the DWP dataset, some pilots reported that records appeared more than once with differing match scores, which was a function of the algorithm producing matches based on address only as well as address and name.

Going forward, there will be further developments of the algorithms used and efforts to standardise the data (for example through the inclusion of a UPRN) which

[10] Commonly reported issues with matching on names included: interchangeability of first names and middle names; use of initials for middlenames; name changes following change in marital status not being updated; and use of shortened versions of names (e.g. Elizabeth to Liz). Commonly reported issues with matching on address included: differences in numbering/naming of properties (particularly in relation to flats and shared accommodation).

[11] This occurred as a result of the different data sources being matched by different organisations (e.g. DWP and MoD matched the data themselves, whereas the DfE/BIS/SLC data was sent to the CO to be matched by an IBM consultant)

should mitigate against the issue of duplication within the records. Nevertheless, it should be noted that this caused issues for areas involved in this pilot.

Key finding: Further development of the algorithms used for data matching is required to reduce the occurrence of duplication/inaccuracies within the data.

As a result of the issues described above the pilot sites reported having to invest significant amounts of time and resource in cleansing and preparing the data before they could reach a point where they could begin analysing the data and carrying out any additional matching with local data sets. This posed particular issues as areas had not anticipated the amount of work that would be required to do this and many registration teams reported that they did not have staff with the relevant IT skills required to complete the task (resourcing requirements are discussed in further detail later in the report).

Another issue commonly raised in the feedback from LAs related to their knowledge and/or understanding of the data provided to them. Many of the areas reported that they did not feel equipped with enough detail on the data to be able to effectively interpret it, as highlighted by one of the interviewees:

> *"we had no key to what that match quality meant. And we also had no idea of the time of the data; how old it was. Was it six months, was it a year, was it two years? There was no qualitative, sort of, information"*

A number of areas reported that having more direct contact with DHOs could have assisted with this:

> *"I think it could have been easier if there'd maybe been some sort of technical workshops or something like that, where the actual owners of the data were sitting round the table ... They could, you know, quite clearly and concisely give you detail about how they hold their data and what it means to them and, you know, technical people can speak at that level to try and understand, give you that head-start so you know what you're looking at"*

Feedback from the DHOs similarly suggests that more direct contact between EROs and themselves may be beneficial. For example, one suggested that whilst they felt that communications between the Cabinet Office and themselves had been strong, they would have benefited from having a more open dialogue with the ERO's involved in the pilot.

This lack of understanding of the data had a number of consequences. Firstly, it was seen as adding to the time that it took areas to get to grips with and prepare the data for analysis. Secondly it led to many areas reporting that they found it difficult to assess the relative accuracy of the data sets. For example, where one data set placed an individual in a certain address and another placed the individual in a different address they weren't clear which they should assume to be the correct record.

The inclusion of a currency/activity marker was highlighted by many areas as one way of overcoming this problem. Other areas suggested that clearer guidance notes, including on the hierarchy of the data, would be beneficial:

> *"I think one of the things we need to do, as a result of this pilot and part of the learning process is, we need to have a set of rules that we could set questions, rules for everyone to call,*

that we go through in deciding whether external data should be used."

Key findings:

LAs would benefit from clearer guidance in relation to the data provided to them, including a description of the data and the data fields and a rough breakdown of the numbers of records they are likely to receive. This guidance should also include a description and explanation of the matching process and match rate scores.

Greater direct contact between the DHOs and the LAs may be beneficial in resolving queries and helping LAs to interpret the data sent to them.

Other suggestions for how to improve the quality of the data

Overall, the feedback from the pilot areas highlighted a need to simplify the data received by electoral administration teams and to make it more accessible. In addition to the points raised in the paragraphs above, a number of areas fed back that the process could have been improved by producing data in a format that would be compatible with the EMS databases used by areas to store and manage their electoral registers.

Whilst the design of the pilot focussed on testing the usefulness of the data rather than the model of the final computer system, the importance of compatibility between systems was recognised and EMS suppliers were provided with details of the pilot ahead of the data matching (e.g. data specifications). Nevertheless, a number of areas had difficulties in uploading the data they received onto their EMS systems which they felt would have been an easier way of managing the data. It should however be noted that this was not an issue across all pilots/EMS systems and some areas were able to upload data onto the EMS with relative ease[12]. In addition, where upload difficulties were encountered these were generally resolved after help from Cabinet Office or the EMS supplier.

Another way in which LAs reported that they felt the data could be simplified was by providing separate data files for records that were matched and those with no matches/'fuzzy' matches. It was also suggested that less information could be provided for the matched data, or that only unmatched data or 'fuzzy' matches could be returned to the LAs enabling them to focus on progressing these particular cases.

Local areas also reported that a small but significant proportion of the mismatches identified in the data related to the inclusion of individuals who are ineligible to vote, primarily owing to their nationality. Many of the areas suggested that, if possible, the inclusion of nationality information within the data would be beneficial in identifying these individuals more easily as currently they were reliant on cross referencing the data against locally held information.

Key findings:

LAs would benefit from receiving data in a more simple and accessible format, additional suggestions for how this could be achieved include:

a) Providing separate data files for records that are matched and those with no matches/'fuzzy' matches, or only providing detailed data on mismatches

[12] Different LAs have different systems, supplied by different organisations.

b) Providing data in a format that is compatible with all existing EMS databases, and maintaining an ongoing dialogue with EMS suppliers who offer valuable input regarding problem resolution.

LAs reported that the inclusion of nationality information in the datasets, if available, would assist LAs in identifying the records of individuals who are ineligible for inclusion on the register.

Resources

The resource required to process and analyse the data varied across local areas, in part owing to the different number/type of data sets matched against and the volume of records/size of the area being matched. However, there were some common themes that can be drawn out from the feedback from pilot areas.

Firstly, many areas emphasised the need for specialist IT resource within electoral administration teams in order to process the data.

> *“"one of the key lessons as a project for us, is that, you know, I think officers are going to need IT support for dealing with this data"*
>
> *"I fear for people who don't have the back-up of ICT data departments. If we were just trying to use an Excel sheet, I think it would be very difficult. "*

Many of the local areas were unable to access this resource from within their team and therefore had to buy this resource in for the pilot scheme.

In addition, the issues experienced in relation to the format/incompatibility of the data described earlier resulted in many areas processing/analysing the data manually. The additional time required to review records individually was reported to have significantly increased the resource required to undertake the matching, which many felt would not be sustainable going forward.

These points highlight a potential capability (and cost) issue for any future roll-out of data matching. However, it should be noted that in many cases this resource was concentrated on the technical aspects of unravelling file formats and cleansing or preparing the data to get it in a format that electoral administration teams could then use for the purposes of follow-up. Therefore, if future developments of the data are successful in producing a more accessible data set, the required level of involvement of IT expertise may be reduced. Nevertheless, many areas felt that electoral administration teams would need further training to be able to process the data – for example advanced Microsoft Excel training.

Key finding: Processing and analysing the matched data requires advanced IT skills, which was identified as a current skills gap within many electoral administration teams.

Public feedback

Some pilots chose to pro-actively advertise the pilot (e.g. press releases) whilst others opted not to. Whilst data on public feedback was not systematically captured pilot areas were asked to report on this within the qualitative element of the evaluation. The majority of areas reported very little feedback from members of the public regarding the use of their data for the purposes of the pilot. In

the minority of incidents where individuals were reported to have raised queries these tended to relate to concerns about individuals being incorrectly identified as residing at their address (and any potential implication of fraudulent activity occurring) or feeling "suspicious" about why their details were being checked. However, in most cases areas reported that providing further explanation of the purposes of the pilot and how the data was being used satisfactorily addressed these concerns.

Key finding: Generally, the level of public interest or concern regarding the pilots was reported to be low.

Follow up work

Each pilot area adopted a different approach to their follow up work, which often included some form of local matching[13] and then sending out letters and/or canvassing households or individuals who had been identified as potentially missing from the register. Details of the approaches that individual areas adopted are included in chapter two, however it should be noted that these approaches may not reflect the original proposals put forward by the pilots. Many areas had to change or adapt their methodology in some way due to delays in starting the pilots or differences in the way in which data matching was conducted compared to how they originally envisaged (e.g. believing they would contain UPRNs or that the matching could be done locally).

The biggest factor was inevitably the timing of the canvass and the resources that were therefore available to undertake the pilot work, as well as the desire to differentiate the two activities for the purposes of the evaluation and to avoid confusing the public. The level of confidence the LAs had in the data also influenced their approach and their willingness to approach members of the public based on the data, particularly in relation to individuals identified as potentially being on the register when they should not (i.e. electoral fraud). Where LAs had carried out local matching they tended to report having more confidence in this data, because they were able to see the detail behind it and have a better understanding of its strengths and limitations. Some respondents suggested other local sources that they believed could be useful if they were able to access them such as local education records and health data.

[13] Many areas reported carrying out local data matching, where the electoral register data is compared to locally held data sources, as a standard part of their work to maintain the register. A number of these incorporated this matching as part of the follow up work undertaken for the pilot data, which in most cases involved checking the data against council tax records, although some other records were also reportedly used, including for example Customer Record Management databases and Housing databases.

CHAPTER 4 – THE POTENTIAL IMPACT OF DATA MATCHING ON THE ELECTORAL REGISTER

This section of the report explores in more detail the results of the data matching, subsequent follow up activities and stakeholders views on the effectiveness of data matching for the future.

4.1: Contextualising the data

The original purpose of the data matching pilot was to assess whether public administration data sets could be used to identify missing electors (particularly traditionally under-registered groups). However, the ability to test this was limited by a number of issues, which it is important to note when considering the findings presented in this chapter:

- Owing to the timetable slippage the analysis and follow up work undertaken by the pilot sites took place at the same time as the annual canvass. As a result, where individuals have been added to the register, it has been difficult to disaggregate the impact of data matching from the standard canvass activities.
- The data sets obtained in the pilot contained limited demographic details. As a result where missing electors have been identified through data matching it has not been possible to carry out a robust assessment of who these individuals are i.e. whether they belong to those traditionally under-represented groups.
- As detailed in chapter one, the pilots adopted different approaches to the follow up making comparisons between them difficult. In addition, in order to inform the evaluation, each pilot site was requested to return monitoring forms to the EC and the Cabinet Office detailing the results of their analysis of the data matching and the outcomes of their follow up activity. Within these forms a number of inconsistencies were observed in relation to the interpretation of the variable fields and the differing match rate thresholds applied across areas[14].

Whilst these factors mean that it has not been possible to robustly assess the effectiveness of data matching for the purposes of the identifying missing electors, the pilot has provided an important opportunity to test the data and, as discussed in the previous chapter, a number of ways in which the process could be improved for future exercises have been identified.

In addition, in the course of the pilot an alternative use for data matching has been identified, namely as a tool for pre-verification for the purposes of the introduction of individual electoral registration. The results of the data matching pilot have shown that a

[14] The match rate threshold is the match score at which it is assumed that records have been accurately matched. Pilot areas were able to set this threshold themselves and these differed between areas.

high proportion of individuals currently on the electoral register could be matched within other public data sets. Using data matching as a mechanism for verifying these individuals' details offers the opportunity to 'passport' a high proportion of the electorate across to the updated registers without requiring them to individually produce personal identifiers (a requirement of IER).

The Government Response to pre-legislative scrutiny and public consultation on IER published in February 2012[15] outlines the intention to simplify the transition to IER for many electors through the use of such pre-verification. The findings presented in this chapter therefore consider the effectiveness of the data matching both in relation to identifying missing electors and for the purposes of pre-verification.

4.2: Findings from the DWP and 'hub' data sets

This section of the report explores the initial results of the data matching and then the results of the pilot follow up activity for the DWP and the 'hub' data sets. The 'hub' datasets include BIS, DfE, DVLA, HEFCE and SLC data, all of which were centrally matched in the Cabinet Office data hub specifically created for the pilot. Findings from the other data sets (Royal Mail, MoD and the local 'Citizens Account' data matched by one pilot site) are presented separately in Section 4.3 owing to the difference in the data sets and their application within the pilot.

Initial results of the data matching - DWP

The DWP data set was the data set that was used by the greatest number of pilot sites (18) and covered the widest range of population groups. As discussed in the previous chapter, a number of issues regarding the volume and quality of data supplied to the LAs were experienced. Following initial feedback from the pilot areas DWP revisited their data and were able to make some refinements to the matching process. The key changes to the process were:

a) a currency limit of 3 months was applied to records that were found on the DWP data set but not on the electoral register
b) some minor amendments were made to the match scoring
c) efforts were made to remove duplicate ERO and duplicate CIS records

As a result of these changes the volume of additional records which were found in the DWP data set, but not on the ER, reduced by an average of 70 per cent[16]. The data presented below is based on this second set of data and therefore differs from the data used by local areas for the pilot[17]. However, as the data represents an enhancement of the original data used by the pilot sites it can be seen to present a more accurate picture of the potential of the DWP data set. In addition, it was possible to apply a consistent threshold for assuming a match to this data meaning that it is possible to more accurately observe the differences in match rates across areas.[18] In this data a strong match is where the first name, surname plus the postcode and/or first line of address were the same in both datasets.

[15] The full response can be accessed at http://www.cabinetoffice.gov.uk/resource-library/ier-command-paper

[16] This is based on data for the 5 out of 9 pilot areas who were interested in carrying out further matching against this data and matched against their whole register. In these areas the volume of records returned which were in the DWP data but not on the ER fell from an equivalent of nine per cent of the electoral register (on average) to an equivalent of two percent of the electoral register (on average)

[17] The second phase of data was commissioned once the pilot had begun and therefore it was not possible for many of the pilot sites to process this additional data owing to time and resource constraints.

[18] During the matching process each record matched to the DWP data was assigned a match score between zero and one hundred per cent. For the DWP data presented here any record achieving a zero to 20 per cent score is counted as a 'no match', over 20 per cent but less than 65 per cent a 'weak match', whilst any record with a score of 65 per cent or above is considered as a 'strong match' and therefore accepted as a match. In practice, as part of the pilot processes individual areas selected their own threshold for accepting a record as a match and this varied between areas making subsequent comparisons across areas challenging.

Table 4.2a: Summary of match rates of electoral register data against DWP data

Area	Full / partial match of register	% No Match	% Weak Matches	% Strong Matches	Total records matched
Camden	W	43%	4%	53%	173,346
Colchester	W	27%	3%	70%	136,926
Forest Heath	W	32%	3%	65%	45,695
Greenwich	W	30%	4%	65%	183,784
Lothian	W	28%	2%	70%	654,515
Newham	W	44%	7%	50%	216,680
Stratford	W	21%	2%	77%	100,942
Tower Hamlets	W	44%	6%	50%	189,661
Wigan	W	19%	3%	78%	250,708
Wolverhampton	W	24%	4%	72%	192,738
Blackpool	P	35%	3%	62%	37,236
Forest of Dean	P	60%	0%	40%	1,143
Glasgow	P	55%	2%	43%	47,660
Manchester	P	39%	5%	57%	17,692
Peterborough	P	34%	9%	57%	8,098
Southwark	P	37%	6%	57%	30,758
Sunderland	P	39%	3%	57%	9,311
Teignbridge	P	22%	2%	75%	33,934

The above table summarises the match scores achieved comparing electoral register data with DWP data for each of the relevant pilot sites. As would be expected, the average match rate (i.e. the proportion of the records on the electoral register with a strong match to a record within the DWP data set) is lower amongst those areas that only matched part of their register (55%) compared to those that matched the full register, which is likely to be a result of sampling bias[19].

Amongst those areas that matched their whole register, on average two-thirds (66%) of records could be strongly matched within the DWP data (range 50 -78 per cent). The average match rate was higher in areas outside of London (73%; range 65%-78%) compared to London areas (54%; range 50%-65%), which may be a feature of the comparatively high population churn associated with large urban/metropolitan areas. This highlights the potential for the use of the DWP data set as a tool for pre-verification for the purposes of individual electoral registration.

It should however be noted that even where a record is successfully matched there remains the potential for dual inaccuracies (i.e. where both sets of data have inaccurate information). It has not been possible to accurately assess the level of dual inaccuracies across the data sets in this pilot, however data collected in one pilot site (Colchester) indicates that this is not likely to be a significant issue. Colchester checked the matched records form the DWP database against locally held records and found dual inaccuracies in less than one per cent of cases. Further exploration of this issue in other areas and other date sets would however be beneficial to provide greater certainty in relation to this finding.

[19] This is because local areas were more likely to select the areas which traditionally had relatively lower registration rates among the general population and/or higher than average proportions of traditionally under-registered groups such as students .

Initial results of the data matching - 'Hub' data (BIS, DfE, DVLA, SLC and HEFCE)

As detailed in chapter two, owing to data the data security requirements of transferring personal data, DHOs would only agree to the matching of the data sets if it was undertaken centrally, with only exception reports being released to local areas. Whilst MoD, DWP and Royal Mail had the capability and/or capacity to undertake the matching themselves the other DHOs were unable to facilitate this and as a result the matching of the BIS, DfE, DVLA, SLC and HEFCE data took place at the Cabinet Office, carried out by a data matching specialist from IBM.

As a result, the matching process differed from that used for the DWP data sets owing to the application of different matching algorithms. For the 'hub' data each record was either marked as unmatched or, where a match was identified, given a score ranging from 81 to 118 to indicate the relative strength of the match. The matching algorithm used in this process was very sophisticated but (unlike the DWP algorithm) tended only to identify strong matches in the great majority of cases. Whilst this is very useful for confirmation/verification, feedback from LAs suggested that they appreciated being able to see more examples of 'fuzzy' matches - and experience showed it was the DWP algorithm which enabled this best.

The data presented below is taken from this central hub and therefore, like the DWP data presented above, may differ from the data provided by the pilot areas in their end of pilot reports

Table 4.2b provides an overview of the match rates for each of the 'hub' data sets in each pilot site. As with the DWP data the match rates for areas who only included part of the register are more likely to differ from other areas owing to the specific population groups they cover. For example Forest of Dean has considerably higher match rates that other areas but this may be expected as they only included entries for attainers (16./17 yr olds) who, given the nature of the data sets are more likely to be included within them (see chapter two for an overview of the data sets).

Table 4.2b: Proportion of records matched within the 'Hub' data sets by pilot area

Area	Whole or partial match of register	BIS	DFE	DVLA	SLC	HEFCE
Blackpool	P	6.3%	0.1%			
Camden	W	4.1%	0.5%		1.2%	5.4%
Colchester	W				1.5%	
Forest of Dean	P	56.9%	36.8%	66.8%		0.6%
Glasgow	P			28.1%	1.4%	
Greenwich	W	4.7%	0.9%	51.5%		4.1%
Sunderland	P	6.3%	0.2%		1.8%	6.6%
Teignbridge	P			61.7%		
Tower Hamlets	W	4.7%	0.5%			4.8%
Wigan	W			67.8%		
Wolverhampton	W		0.0%			3.7%

Comparing the average match rates of only those areas that matched the whole register shows that the DVLA data set has a considerably higher average match rate than the other data sets (60 per cent). The BIS data set has the next highest average rate (4.5 per cent) with DfE and SLC match rates the lowest, averaging at around one per cent. Whilst only a small number of pilots matched against each of these data sets, meaning that the figures should be treated with a degree of a caution, they do indicate that the DVLA data set has the potential to be a useful tool for the verification of records on the register given its relatively high match rates.

The Forest of Dean example is also interesting, as the relatively high match rates observed could indicate that other data sets, notably BIS, may be particularly useful for targeting attainers. However, further piloting with additional areas would be needed before any firm conclusions on this could be made.

Key findings:

In addition to DWP, data collated in the pilot suggests that the DVLA data set also demonstrates comparatively high average match rates and could therefore be a useful tool for verification of the register.

Whilst the BIS, DfE and SLC data sets have much lower match rates, there is some evidence to suggest that they (BIS in particular) may be beneficial in identifying specific groups who are traditionally under-registered, namely attainers. However, further piloting with additional areas is required to test this assumption.

The data described above illustrates the match rates for each of the data sources individually. As the hub data sets were matched together, unlike the other data sets, it is also possible to use this data to explore the potential impact of matching against a combination of data sets. Table 4.2c overleaf presents the combined match rates of the pilot areas who matched against more than one of the BIS, DfE, DVLA and SLC data sets.

As would be expected, given the findings highlighted earlier, the table shows that those areas which matched against the DVLA database successfully matched a much greater proportion of their electoral register records than other areas. However the data also provides a useful illustration regarding the potential use of the data sets to identify individuals who are not on the register.

Comparing the total number of records from all data sets to the total records from the electoral register shows us that, with the exception of the Forest of Dean, a number of additional records were found in the data sets that were not included on the register. In some cases these records may represent additional voters, however the feedback from the pilots (as explored in chapter three) suggests that issues concerning the quality of the data mean that this cannot be assumed in many cases.

One way in which it is possible to have greater confidence in the data is where an individual appears on more than one of the data sets matched. The data shows that whilst a number of these records were identified in the pilot areas, that number is relatively small. Across the pilot areas matching against the 'hub' data sets, only between zero and two percent of those records included in other data sets but not the electoral register data were found in more than one data set, which represents less than one per cent of the total electoral register entries matched.

Key findings: The pilot data suggests that using data matching to more accurately identify potential new registrations through matching those records that do not appear on the ER across the other (non ER) data sets offers minimal benefit. However the extent to which the quality of the data may have impacted on this finding is unclear.

Table 4.2c: Proportion of records matched across all data sets (Electoral Register (ER), BIS, DfE, DVLA, SLC)

	Datasets matched				No. of records matched			Match Results		
	BIS	DfE	DVLA	SLC	Total records from ER	Whole or partial match of register	Total records from all data sets	Individuals on ER matched to at least one other data set	Individuals not on ER but matched within at least two other data sets	Proportion ER records matched within the data sets
Forest of Dean	√	√	√		747	P	747	746	-	99.9%
Wigan			√		241,408	W	321,288	159,495	-	66.1%
Teignbridge			√		33,413	P	46,605	20,358	-	60.9%
Greenwich	√	√	√		180,424	W	288,196	94,960	2,455	52.6%
Glasgow			√	√	46,974	P	66,459	13,276	81	28.3%
Sunderland	√	√		√	9,185	P	10,002	721	2	7.8%
Blackpool	√	√			36,918	P	39,382	2,284	3	6.2%
Camden	√	√		√	172,261	W	179,875	9,610	37	5.6%
Tower Hamlets	√	√			187,797	W	198,104	9,375	45	5.0%
Colchester				√	135,605	W	136,841	1,867	-	1.4%

Findings from the follow up (DWP, BIS, DfE, DVLA and SLC[20])

This section of the report presents the findings in relation to how LAs were able to use the matched data that they received, including whether the data was used and if so was effective at:

a) identifying people who were not on the register but were entitled to be; and
b) identifying people who are currently on the register when they are not entitled to be.

In order to help answer these questions pilots were encouraged by the Cabinet Office and EC to have a control group to help assess the effectiveness of data matching compared to the canvass in adding and deleting people from the electoral register. Most did adopt this approach in some form or another, with a control group randomly selected and then canvassed in the normal manner. For some areas or some data sources, where the sample sizes were smaller, this did not always occur and therefore it is more difficult to assess the impact of the data compared to the canvass.

The data presented below (pages 44-47) are taken from the pilot data monitoring forms completed by the pilot sites at the end of the pilot. As noted previously, each pilot site adopted a different approach to the pilot, including follow up activities, and there were some discrepancies across areas with regard to how they interpreted the data fields in the end of pilot report. Therefore the results should be treated with a degree of caution, particularly when looking across areas[21].

Identifying people who were not on the register but were entitled to be

Table 4.2d summarises the results of the follow up work undertaken by LAs with individuals who were identified on the DWP data set but not the electoral register. It shows that of those pilot sites that included a control group, in all but one area (Southwark), a greater proportion of individuals were added to the register in the control group than in the data matching group.

Tables 4.2e-4.2i summarise the results the results of the follow up work undertaken by LAs with individuals who were identified on the 'hub' data sets (i.e. BIS,Dfe, DVLA & SLC) but not the electoral register. Tables 4.2e-h show the results for the individual data sets, however as the hub data sets were matched together centrally some areas did not distinguish between the data sets for the purposes of follow up and therefore their results are presented as combined in table 4.2i. These tables also show that, of those pilot sites that included a control group, across all areas, a greater proportion of individuals were added to the register in the control group than in the data matching group.

As detailed in chapter three, a number of areas also matched their data against locally held data sets as part of their follow up activities. Only a limited amount of data was provided on this local matching meaning that it is not possible to draw any firm conclusions on the effectiveness of such local matching. The information available does indicate that there is some potential to use local datasets as part of the data matching process. For example, Southwark reported that of those people appearing on the DWP data but not on the electoral register 14 per cent appeared on at least one other locally held dataset. Of these records however, 88 per cent were added to the register during the annual canvass or by rolling registration between data transfer dates.

[20] The HEFCE data is not included here as the data set did not include complete addresses and therefore could only be used for the purposes of initial matching/verification.
[21] In addition, as highlighted previously, the numbers will not necessarily match with data from the central matching presented in earlier sections of the report

Overall, the results of the follow up activities undertaken by pilots do not provide any evidence that data matching is a more effective mechanism for identifying and adding missing electors than the annual canvass. It is difficult to know with any certainty why this is. The currency of the data may have been a factor, with many areas reporting that a significant proportion of individuals they followed up as a result of the matching had simply moved on from the address. It is possible that the timing of the pilots may have had an impact, as individuals who may otherwise have been captured through the data matching were instead picked up in the annual canvass. For example Southwark began their data matching and follow up work at an earlier stage than other areas and found that the proportion of individuals who were added to the register through the DWP data matching and in the control group were very similar[22]. However, it is not possible to draw any firm conclusions from the experience of one area and so further piloting of data matching outside of the canvass period would be needed to test this further.

In addition, the small sample sizes and the differing approaches to the follow-up work mean that it is difficult to accurately assess any trends in relation to the effectiveness of the data sets in relation to targeting specific groups (e.g. students). It would therefore be beneficial to conduct further piloting where this can be assessed more completely by applying a more consistent methodology across pilot areas.

Key findings:

Data matching did identify some missing electors, however the results of the pilots' follow-up work suggests that it is a less effective means of identifying and adding missing electors than the annual canvass. It is not clear the extent to which issues with the currency of the data matched and the timing of the activities impacted on these results and so further research is required to confirm this.

Interpreting the results of the pilots follow up work is further complicated by the fact that the pilots adopted varying approaches to this work, and where specific groups were targeted (e.g. students) involved small sample sizes.

As a result of the above, further piloting, undertaken at a different time of the year and incorporating a more consistent methodology across pilots, would be beneficial.

[22] This meant that some follow up letters were sent in advance of the canvass related communication, although there remained some overlap between the data matching and the canvass related activities.

Table 4.2d: Results of follow up of people identified in the DWP data set but not the electoral register

		Numbers followed up		People added to the register				People removed from the register			
	Pilot	Number to follow up	Number retained for control group	Total added to register from follow up	Total added to register from control group	% added to register via data matching	% added to register via control group	Total deleted from register from follow up	Total deleted from register from control group	% deleted from register via follow up	% deleted from register via control group
Control group included	Blackpool[a)]	2,466	2,466	727	771	29%	31%	54	51	2%	2%
	Camden	9,230	1,234	387	197	4%	16%	0	0	0%	0%
	Colchester	1,423	1,677	74	721	5%	43%	0	0	0%	0%
	Forest Heath[b)]	4,696	200	1398	not stated	30%	not stated	not stated	not stated	not stated	not stated
	Forest of Dean	33	70	5	51	15%	73%	0	0	0%	0%
	Greenwich[c)]	3,713	4,176	211	543	6%	13%	n/a	n/a	n/a	n/a
	Lothian[d)]	10,215	9,875	1,139	3,061	11%	31%	n/a	n/a	n/a	n/a
	Southwark	5,829	648	2,545	272	44%	42%	2,173	0	32[e)]	0%
	Stratford-upon-Avon	1,035	141	10	103	1%	73%	0	0	0%	0%
	Wigan	5,012	1,138	187	307	4%	27%	not stated	not stated	not stated	not stated
	Wolverhampton	3,868	6,992	723	886	19%	13%	not stated	not stated	not stated	not stated

..continued overleaf

Table 4.2d continued

	Pilot	Number to follow up	Number retained for control group	Total added to register from follow up	Total added to register from control group	% added to register via data matching	% added to register via control group	Total deleted from register from follow up	Total deleted from register from control group	% deleted from registe r via follow up	% deleted from register via control group
No control group included	**Glasgow**	331	0	94	n/a	28%	n/a	67	n/a	20%	n/a
	Manchester	no follow up	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	Newham	1,902	0	79	n/a	4%	n/a	n/a	n/a	n/a	n/a
	Peterborough	no follow up	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	Sunderland	2,408	0	297	n/a	12%	n/a	0	n/a	n/a	n/a
	Teignbridge	no follow up	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	Tower Hamlets	no follow up	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Notes: **a)** Blackpool figure for number of people to follow up and control relate to properties rather than individuals and are taken from their evaluation report **b).** Forest Heath figures on electors added may include individuals who also received a canvass form **c**) For consistency all data in the table as included is reported as included in the end of pilot reports (except Blackpool, see point a), however Greenwich note that the figures on people added to the register exclude those added through the usual canvass activities (an additional 943 in the data matching group and an additional 1,113 in the control group) **d)** Lothian focused on void properties where a possible elector had been identified, but also did some mismatches and these figures are for the different activities taken together.

Table 4.2e: Results of follow up of people identified in the DfE data set but not the electoral register

Pilot	Number to follow up	Number retained for control group	Total added to register from follow up	Total added to register from control group	% added to register via data matching	% added to register via control group
Blackpool	2,466	2,466	727	771	29	31
Forest of Dean	12	44	4	35	33	80
Greenwich	244	391	18	141	7	36
Wolverhampton	560	no control but found potential of 4,022 electors	331	no control but 1,038 added via canvass	59	n/a

Table 4.2f: Results of follow up of people identified in the SLC data set but not the electoral register

Pilot	Number to follow up	Number retained for control group	Total added to register from follow up	Total added to register from control group	% added to register via data matching	% added to register via control group
Colchester	39	no control	2	not stated	5	n/a

Table 4.2g: Results of follow up of people identified in the BIS data set but not the electoral register

Pilot	Number to follow up	Number retained for control group	Total added to register from follow up	Total added to register from control group	% added to register via data matching	% added to register via control group
Forest of Dean	31	75	6	73	19	97
Greenwich	724	849	24	136	3	16

Table 4.2h: Results of follow up of people identified in the DVLA data set but not the electoral register

Pilot	Number to follow up	Number retained for control group	Total added to register from follow up	Total added to register from control group	% added to register via data matching	% added to register via control group
Forest of Dean	34	94	9	n/a	26	n/a
Greenwich	3,399	3,505	38	140	1	4
Teignbridge	none undertaken	n/a	n/a	n/a	n/a	n/a
Wigan	1,701	0	420	n/a	25	n/a

Table 4.2i: Results of follow up of people identified in combined 'hub' data sets but not the electoral register

Pilot	Number to follow up	Number retained for control group	Total added to register from follow up	Total added to register from control group	% added to register via data matching	% added to register via control group
Camden[a]	383	0	17	n/a but 49 of follow up registered via canvass	4	n/a
Glasgow[b]	102	66	3	not stated	3	n/a
Greenwich[c]	247	17	4	4	2	24

Notes: **a)** Camden data is for BIS, DfE and SLC combined **b)** Glasgow data is for DVLA and SLC combined, Glasgow also deleted 14 people from the register as a result of follow-up work (14%) **c)** Greenwich data is for some BIS, DfE, and DVLA data combined.

Identifying people who are currently on the register when they are not entitled to be

A small minority of pilot areas (3) reported removing anyone from the register as a result of the data matching. Where this occurred it was predominantly because the individual was discovered to have left the property as opposed to there having been an indication of fraudulent registration. Only one of these areas (Blackpool) included a control group in their pilot and, like their results for the proportions of additions to the register, found similar proportions of electors were removed in both the pilot group and the control group.

Findings from the qualitative interviews with LAs suggest that the majority of areas had not intended to or were not confident enough to use the data for this purpose. A key theme that arose in the responses was that there was not enough trust in the data to feel happy challenging an individual on the basis of a mismatch:

> *"We wouldn't take anybody off the register, just because we got data matching saying somebody else was living at that property, because who's to say they haven't moved into that property? And that person has the right to remain on that register until he asks to come off that register"*

> *"I wouldn't write to anybody just on the national data alone because I don't think it's reliable...We weren't happy to write to those people on the basis of what we were looking at."*

Some areas also reported that they did not feel that fraudulent registration was a significant issue in their area and therefore could not justify the necessary allocation of resources to this element of the follow up work.

Key finding: The majority of pilot areas did not use the data matching as a mechanism for removing electors from the register. Feedback from the pilot areas suggested that this was because they did not feel confident and/or justified in using the data matching for this purpose, or that they didn't perceive that fraudulent registrations were an issue in their area.

4.3 - Findings from the other data sets (Royal Mail, MoD and Citizens Account)

Royal Mail

This data set was included in the pilot as a potential source of information for identifying recent movers, one of the groups that is traditionally under-represented on the electoral register.

Royal Mail provides a redirection service to members of the public who wish to have mail which is addressed to them forwarded to a new address. The Redirection application is verified at point of application.

The National Change Of Address – NCOA Update and NCOA Suppress files are taken from the Royal Mail Redirections database. NCOA Update contains the names and both the new and old addresses of residential customers who have taken out a permanent redirection, and the NCOA Suppress file the moved from address only.

NCOA Update data is made up from only those customers who have applied for a Redirection, and address details are only provided where customers have provided the relevant permission for Royal Mail to share their data, therefore it only includes data from a small section of the population, and can be used for update purpose only, not

suppression. The NCOA Suppress data is made up from expired Redirections and can be used for suppression purposes only.

The length of time required to agree data sharing protocols with Royal Mail was greater than the majority of other data sets, owing to the sensitivities and restrictions of matching a commercial database, terms and conditions of use for the NCOA Update and NCOA Suppress databases, and that Royal Mail is governed by RIPA which meant additional governance had to be put in place with not only the signature of an End User License but also a Public Body End User Agreement.

As a result only one pilot site was able to provide data from their electoral register to be matched against, and the matched data could not be returned to the pilot site in sufficient time to enable any follow up work undertaken. The findings of this matching do offer some indication of the potential for data matching using Royal Mail, however further piloting is required to test this in practice, and across a wider area.

Approximately 137,000 electoral register records were matched against the Royal Mail NCOA database, of which approximately four per cent (5,225) were matched within the database.

Of those four per cent, approximately 42 per cent had moved within the area covered by the ERO, 31 per cent had moved outside of the area and the remainder (27 per cent) had no forwarding address.

A recent study on the completeness and accuracy of the electoral register suggests that the completeness of the register declines by an average of ten percentage points in a year, owing mainly to population movement (Electoral Commission, 2012). This suggests that the Royal Mail data set may have the potential to capture a reasonable proportion of these, although we do not know what proportion of these recent movers would be picked up through the usual canvassing activities.

Key finding: Findings from the pilot suggest that Royal Mail data has the potential to identify a proportion of recent home movers, a group which has been traditionally under-registered. However, given the limited opportunity to test the data in the current pilot, further piloting of this data set is necessary to provide greater certainty of the relative benefits of using this dataset.

Ministry of Defence (MoD) data

Maintaining the registration of service personnel presents some particular challenges for electoral administration teams. Service personnel are eligible to register as an ordinary voter or an overseas voter, but are also eligible to register as service voters by way of a service declaration. This option is open to all personnel (including spouses or civil partners) and is seen as particularly suitable for personnel posted overseas or likely to be posted abroad in the near future. Following the introduction of The Service Voters' Registration Period Order 2010, the Service declaration period was extended to five years, to help ease the burden on Service personnel, and ensure that they remained on the register and able to vote. This can however make it difficult for electoral administration teams to maintain the accuracy of this section of the register as they may be less likely to receive notification on when service personnel move in and out of their areas. Therefore data matching with MoD presented a useful opportunity for EROs to check the completeness and accuracy of this section of the register.

Six local areas opted to use the pilot as an opportunity to explore registration of service personnel, although only four were able to use the data in practice. The MoD agreed to match the electoral register data of these areas against two of their datasets. The first of these data sets focussed on properties (using a database known as the ANITE housing system), looking at whether the properties marked on the electoral register as military properties were also on the property lists held by the MoD and highlighting any discrepancies between the two.

The results of this matching are illustrated in table 4.3a below. The small number of pilot sites using this data means that caution should be applied when interpreting this data, however the findings show that in three out of the four areas there was a very high level of consistency between the properties identified as military owned on the electoral register and the MoD property data. Three of the four areas provided data on the follow-up work undertaken on the basis of the data, none of whom reported identifying any new properties that weren't previously known to them, although a small number of properties which had been included on the ER but not as service properties were identified.

The second data set included personnel records collected from the MoD Joint Personnel Administration (JPA), enabling individuals to be matched against the electoral register. The matching was conducted by the MoD, however for reasons of data security they were only able to provide basic information on whether or not the electoral register data was matched within this database. In addition, due to the relative sensitivity of this data compared with the property data only three of the pilot sites were able to obtain this data.

The results of this matching are shown in table 4.3b (overleaf). It shows that the match rates for the data varied between areas. The lack of additional information in the data meant that it was not possible to make any assessment of whether this is indicative of inaccuracies within the electoral register or within the dataset itself. Furthermore, as personnel are responsible for updating their JPA records themselves it was not possible to verify the accuracy/currency of the data, particularly as some of the pilot sites reported hearing of anecdotal evidence that not all service personnel regularly updated their JPA records.

Table 4.3a: Military properties matched between ERO and MoD records

	No. of records provided to MoD for matching	No. of records matched	No. of records not matched (held by ERO not MoD)	Match rate
Rushmoor	1,760	1,748	0	99%
Shropshire	1,169	719	93	62%
Stratford-on-Avon	124	102	20	82%
Wiltshire	5,644	5,471	415	97%

Note: A small number of duplicate records and/or addresses that were outside of the local authority were found in the data, therefore the total records matched and non-matched does not exactly equal the number or records provided for matching.

Table 4.3b: Military service personnel JPA records matched between the electoral register

Pilot	No. of service voters on pre-pilot register	No. of service voter entries matched / confirmed by MoD data	No. of service voter entries not matched by MoD data	No. of reviews of service voters inititated	No. of service voter details amended	No. of electors deleted	Non-response
Rushmoor	500	220	280	280	57	83	140
Shropshire	384	111	273	273	28	34	184
Stratford-on-Avon	128	76	52	0	n/a	n/a	n/a

The data shows that in those areas where reviews of service voter details were undertaken some amendments to the register were made as a result, most commonly to remove individuals from the register where they had moved away. Overall however, a key theme that arose in the interviews with local areas was that whilst data matching had the potential to be beneficial if a greater level of detail could be provided within the data, in their view, the most effective way to drive up registration rates for this group lies in effective engagement with personnel and senior officers. A number of areas cited positive experiences of attending local barracks suggesting that improving registration may therefore be more about direct communication with local military personnel (supported by the central military) than data matching itself.

Key finding: Pilot sites reported finding the MoD data to be of relatively little value owing to the lack of detail contained in the data set. A number of areas suggested that the most effective way to drive up registration rates for this group lies in effective engagement with personnel and senior officers.

Citizens Account Data

Whilst the majority of pilots undertook data matching with national data sets, one pilot area – Renfrewshire – opted to carry out the exercise with a local data set, the "Citizens Account" data. The Citizens Account includes data owned by each of the 32 local authorities in Scotland and managed through the Improvement Service. The Citizens Account is, in effect, an entitlement card giving access to a range of local government, central (Scottish) government and health services. Given the range of services encompassed it was envisaged that the data set would have the potential to identify individuals within groups where registration levels are known to be below average including young people and individuals living in areas with multiple deprivation where the card would give access to social and health care facilities.

Renfrewshire reported being able to confirm that around 30 per cent of the names on the ER were a direct match with the Citizens Account dataset, however they reported that these were largely registrations where recent canvass returns were held suggesting that the data was not as effective as had originally been envisaged at identifying missing electors.

4.4: General views on the future of data matching, including pre-verification

The pilot aimed to explore the potential for data matching to be used as a tool for identifying people who are not currently on the register who should be, with a focus on specific groups who have traditionally been less likely to be registered than others. As discussed in earlier sections of this report, the majority of LAs did not report finding the data matching exercise to be very effective at identifying individuals who are currently missing from the register, primarily as a result of the quality of the date, and the timing of the follow-up work, which clashed with the annual canvass.

Despite this perceived lack of effectiveness many participants could see the potential benefit of data matching for the future, particularly with the advent of individual registration and if the quality of the data can be improved and the timing of the exercise is better:

> *"When it comes to individual registrations, data matching might be better there because you're matching against a name in a house, so, you know, you might have three different names in a house. Yes, it might help you to identify a lot better once you put them in individual registrations"*

> *"I think data matching is going to be something of the future, because everyone is going to be looking at this issue of resources and how to get more verifications on the register without the costs of follow up action, yes."*

> *" I think it could be of use if we have to raise queries on dates of birth and things like that. I'm not really sure whether the process, as it is at the moment, there is any great benefit from [local areas] point of view, because we have quite a good response rate ....I mean, I think I can understand it more if individual registration comes in, and people are able to register on-line, and then automatically, you know, they put in their date of birth and national insurance number or something, and off it goes, and does a match in the background, I can see, yes, that that is very beneficial"*

A number of different suggestions were made for further improving the data matching process. These included: centralising the data matching process and conversely providing the data to the LAs to do the matching themselves; learning from other similar schemes (for example, the DWP Housing Benefit Matching Service); only using local data sets; using other data sets (e.g. credit reference agencies, health, TV licensing, parking services). However there was a lack of consistency between areas regarding these suggestions and most were only mentioned by one/two interviewees.

As highlighted earlier, one of the key findings from the pilot related to the potential for data matching to verify individuals on the register, and therefore to be used as a pre-verification tool for the purposes of the introduction of individual registration. This was echoed in the feedback from a number of the pilot sites as illustrated in the following excerpts from local areas end of pilot reports:

> *"Where the data exchange with the CIS/DWP showed most potential to aid with the transition to IER was in the validation of existing electors. With the proviso that a number of records were either missing in the second exchange or were not possible to reconcile with local property records a significant percentage of those already on the electoral role were identified as present on the CIS/DWP data. Using this to validate an entry on the register (as both current and*

true) without an additional transaction with the elector could allow for a much more targeted and effective transition"

"The single biggest positive to come from the pilots is that (whilst the data did not allow for targeting of missing electors) it did correctly match a high proportion of the settled population. This could be used to avoid potentially unnecessary transactions with the public if it was utilised to allow the automatic transfer over to the new IER register of people about whom a high level of certainty can be achieved. This would then allow for limited resources to be more effectively targeted on encouraging and achieving registrations at those addresses not covered and amongst those potential electors who are currently under-registered and at more risk of falling off the register"

In order to further explore the potential for using data-matching as a tool for pre-verification the Cabinet Office conducted some additional analyses of the data available. The first piece of analysis sought to explore in more detail some of the differences in match rates observed between areas. In order to test the assumption that the differences in the comparative match rates between local areas are likely to be driven, to an extent, by the make-up of the population within them, detailed analysis of the DWP match rates, broken down to Ward level was carried out in two LAs.[23]

By comparing the match rates within a local area, as opposed to across local areas, it is possible to limit the amount of difference between the match rates that may be attributable to differences in approaches to maintaining the electoral register. The analysis revealed that the DWP match rates for data did vary between wards (ranging from 31-80 per cent across one area, and 51-79 per cent in another). The lower match rates were found in those wards that were known to contain relatively high proportions of the population who are traditionally under-registered (e.g. students, individuals residing in temporary accommodation and/or multiple occupancy dwellings) compared to other wards.

These findings should be interpreted with caution, as they are based on a small number of areas. It is also important to note that the lower match rates may arise from a lack of completeness and/or accuracy in either the DWP data set or the electoral register. Nevertheless, this analysis provides further evidence to suggest that data matching could be used as a highly effective tool for pre-verification for a majority of the population. It also highlights the importance of focusing resources on effectively targeting the minority of individuals in the harder to reach groups, which may not be captured by this data.

The second piece of additional analysis sought to explore whether the DWP match rate could be further improved by combining it with the DVLA and other data sets. This analysis explored the additional proportion of the register that could be matched by adding in the additional data sets in sequence. Due to time constraints this analysis could only be carried out in one LA (Greenwich)[24], therefore the results are indicative only, however they showed that in Greenwich, the addition of the DVLA data resulted in a rise in the match rate of just under ten per cent. The inclusion of the BIS and DfE data sets did also lead to further increases in the match rate but these were very small (less than one per cent).

[23] A ward is a division or district of a city or town, used for administrative purposes.

[24] This analysis was conducted by the Cabinet Office. For data security reasons the matching had to be undertaken at the pilot site and it was not possible to conduct similar analyses elsewhere within the legal timeframes of the pilot (after which the data was required to be destoyred)

This suggests that combining data sets has the potential to increase the match rates for the purposes of pre-verification and that, of the data sets tested in the pilot, a combination of DWP and DVLA data is likely to produce the highest match rates.

Key finding: There is some evidence to suggest that combining DWP and other data sets, notably DVLA, has the potential to increase the match rates for the purposes of pre-verification by as much as ten per cent. However as this has only been tested in one area further piloting is required to enable greater confidence in this finding.

4.5: Cost effectiveness

Each missing citizen added to the electoral register cost the equivalent of £50 per person, representing poor cost effectiveness compared to traditional community outreach programmes. This is not reflective of the overall cost effectiveness of using data matching or the pilots however as the primary goal of the pilots was to establish the feasibility of data matching rather than adding citizens to the register.

The pilots successfully demonstrated that 66 per cent of the electoral register could be strongly matched to DWP data and indicated therefore the feasibility of automatically placing matched electors onto the register. This also implies that significant resources could be freed up to target the minority of electors who are not matched (electoral registration officers spent around £47.5m in the 2008-9 canvass period manually collecting information from all electors[25]).

It is less clear from the pilots whether using data matching to find missing people is a feasible option. The pilots did however demonstrate that it was possible to use other sources of public data to find missing citizens and add them to the register to some extent. Future pilots will seek to explore further whether this is the most effective way of adding new citizens to the register.

[25] Electoral Commission, Financial data 08-09

CHAPTER 5 – CONCLUSIONS AND RECOMMENDATIONS

In late 2010 the Cabinet Office invited local registration officers from across Great Britain to put forward proposals for piloting data matching to national datasets not previously available to them as set out in a prospectus (see Annex A). The primary purpose of these pilots was to test the potential benefit of various datasets on the completeness and accuracy of the electoral register, with a view to supporting the move from a system of household to IER in 2014. The 22 pilots who took forward this work were self-selecting and were encouraged to put forward their own ideas and innovations. Whilst these differences in their methodology have made the evaluation more complex, the aim was to allow as many lessons to be learnt or policy and practice as possible. Pilots of this nature had not been undertaken before and as such it was expected that the process and implementation of data matching would not necessarily be a smooth or straight forward experience. However, it was viewed as an opportunity to capture what worked, for whom, in what circumstances in line with a 'realistic evaluation' approach. This report has sought to outline and evaluate both the process of data matching in the pilot areas and the potential impact on the electoral register.

This chapter sets out the key findings and lessons from the previous chapters and draws them together in the form of more definitive conclusions and identifies recommendations for the future.

Process of setting up and running data matching pilots

Before data matching could begin there were significant pieces of work that needed to be completed and in place to ensure legal compliance of sharing the data and that the data was shared in as secure a way as possible. In some instances a lengthy period of time was needed to complete and agree the necessary legal documents between DHOs and LAs. This was a particular issue for Royal Mail data which is subject to commercial sensitivities and for MoD data which is very sensitive and security is of primary concern. The impact of delays in the legal agreements for these datasets was significant as it led to the data being used in a limited way, if at all; by the relevant authorities due to time constraints in the period remaining for the pilots to take place.

Recommendation 1: Adequate time should be allowed for the necessary legal agreements to be in place before any future data matching pilots commence, particularly where the data is viewed as being of a more sensitive nature.

More generally, delays in beginning the pilots led to an overlap with data matching activities and the annual canvass in many pilot areas. This caused numerous problems for pilots in

terms of the resources they had available to complete both pieces of work but also made it more difficult to evaluate the impact of data matching on the completeness and accuracy of the register (this is discussed in more detail below). It was suggested by pilots that any future piloting activity which involved following up missing electors to encourage them to register should take place after the publication of the electoral register in December when it is at its most complete and accurate in order to identify those missing more effectively and assess the accuracy of the register i.e. in January, February or March, or pre-canvass in order to help inform canvass activities i.e. June or July. It should however be noted that data matching for the purposes of pre-verification is likely to be less time sensitive and may usefully be carried out a different time than matching for the purposes of completeness and accuracy, including potentially as part of the canvass activities.

Recommendation 2: Any future data matching piloting activity which requires LAs to conduct additional work needs to be considered in the context of the timing of the annual canvass and the resources available. It may be that it can offer most benefit if conducted pre-canvass in order to inform canvass activities or post canvass to identify those missing electors and check the accuracy of the register. However, some LAs may still find it beneficial to conduct matching (local matching and matching for the purposes of pre-verification in particular) during various stages of canvass activity if they desire.

IT and technical issues relating to the datasets

The pilots were not intended to test the IT capability but nevertheless some useful lessons and issues have been identified and learnt which can help shape the IT development for IER going forward. These have been key lessons for the Cabinet Office which are being embedded in work on digital design and delivery for individual electoral registration. The process of transferring the data was a mixed one for pilots and would benefit from further refinement and testing. Some experienced issues with using secure emails and with their own or DHOs firewalls.

Recommendation 3: Further testing and refinement of transferring data between DHOs and LAs is required to ensure the process runs smoothly.

There were also issues in terms of the consistency of datasets with EMS systems and the formatting of some of the datasets meant that match rates were not always as high as they could be. The key ways in which it was suggested that this could be improved were to standardise address and name formats across national datasets and (and in particular) for DWP to incorporate UPRNs into their data. Improvements could also be made to the matching algorithms to improve matching rates and reduce the number of duplicate records. Again this is a lesson which has been learnt and refined algorithm will be taken forward in any future data matching.

Recommendation 4: Where possible there should be greater consistency

between the national datasets and the electoral register and management system to ensure compatibility. In particular improved standardisation of data formats and the use of UPRNs in national datasets would improve match rates, in addition to more sophisticated algorithms.

One of the biggest issues for pilots was the size of the data files they received back from DHOs, in some cases they were two or three times the size of the their register. A key issue here was the need for the data to be current and accurate and this could have been ensured by only matching to records which had been updated within a shorter time period such as 3, 6 or 12 months. The SI which allowed these pilots to take place did not allow for record dates to be provided and as such pilots were unable to tell how recently the record had been updated (aside from within a two year period) or why it had been updated and therefore make a judgement on the reliability of the information. This is a key learning point for Cabinet Office for any future pilots.

Recommendation 5: Any future data matching should match to records which have been updated or had some activity within the previous 3, 6 or 12 months to ensure they are current and accurate. A record date should be provided and if possible the nature or reasons for the update/activity.

Other suggestions for reducing the scale of the data included only providing mismatches or fuzzy matches or at least providing separate files on matches, mismatches and fuzzy matches – this will be discussed in more detail below when considering the future of data matching. As the file sizes were often very large and sometimes difficult for pilots to understand and analyse, they required quite intensive resourcing and sometimes additional resources were brought into assist as the individuals did not have the relevant technical skills to work with the data. It is hoped that any future matching would reduce these issues by making the data more user friendly and decreasing the scale by improving the matching coding. However, there could still be the need for some specific training or skill sets to be improved within electoral administration teams, for example the ability to use Excel.

More generally pilots expressed a desire for greater guidance and clarity on how to use and interpret the data and detail on how the matching had been conducting and what the match scores meant.

Recommendation 6: Any future data matching pilots should include more detailed guidance on the various datasets; what the variables mean, how they should be interpreted and used and how the matching has occurred. If possible thought should be given to involving relevant EROs and DHOs in the development of methodology at an early stage to ensure greater understanding of the data.

Impact of data matching on completeness and accuracy of the register

The original primary purpose of the pilots was to test the impact of various datasets on the completeness and accuracy of the electoral register, and in particular the ability to find missing electors – especially those traditionally under-registered groups such as students, attainers and service personnel. However, it has been difficult to fully assess the impact due to the overlap with the annual canvass. This overlap has made it difficult to disaggregate the data and assess the counterfactual because even with control groups some individuals received both a follow up letter and registration form as a result of canvassing and a normal household canvass form. Furthermore, there is limited demographic data available on most groups making it difficult to assess whether those who were added to the electoral register were from under-registered groups. Finally, there were also inconsistencies in the reporting data provided to the Cabinet Office and the EC from pilots. These inconsistencies related to differing interpretations of variables and different thresholds for what they did or did not consider to be a 'match'. These factors make it difficult to interpret the data in a robust way and therefore assess the impact on the register in comparison to the annual canvass. However, the pilots have helped to test the data and identified some areas for further possible testing.

The results of the follow up activities undertaken by pilots do not provide any evidence that data matching is a more effective mechanism for identifying and adding missing electors than the annual canvass. However, it is difficult to be certain of the reasons for this:

- the currency of the data (discussed above) could affect responses to data matching follow up as many individuals had moved from the address they were targeted at;
- the timing of the follow up may have been an issue since some individuals may have been picked up by the annual canvass instead – evidence from Southwark where follow up work was completed earlier than in other areas suggests that a similar proportion of people were added from the data matching follow up and the canvass, but it is difficult to draw firm conclusions from just one area (highlighting again the need for further testing to occur at other times of the year); and
- finally there were issues in terms of small sample sizes and varying approaches which make it difficult to interpret the data from DHOs such as BIS, DfE and SLC.

Further testing of BIS, DfE and SLC data is required on a larger scale using as BIS data in particular showed potential for identifying attainers. There was limited opportunity to test the data from Royal Mail on home movers due to delays in the legal agreements, but it does appear to show potential for identifying this key target group who have been identified by recent research (EC, 2011) as being likely to be missing from the electoral register.

Recommendation 7: Further testing of some specific datasets on a larger scale, involving a consistent methodology across pilot sites is needed to see if they can effectively identify missing electors from target groups such as students, attainers and home movers.

The data from the MoD showed a high match rate on addresses but the data on service personnel was limited as the MoD would only confirm if electors on the register remained or

had moved; they did not provide the details of any new electors. The pilots therefore felt that the data was of limited value and that a more effective way of driving up registration rates among service personnel would be to engage with the military at a local level.

The future of data matching

An unexpected potential use and benefit of data matching which has been identified as a result of these pilots is the high match rate between the electoral register and DWP data (on average 66% for those matching the whole register), which means that it could be useful for pre-verifying electors and 'passporting' them across during the transition to individual electoral registration. This will help to ensure that the majority of electors will not have to provide personal identifiers and can be moved straight across, freeing up resources for EROs to focus on the remaining third of electors. More detail on this is set out in the Government's response to pre-legislative scrutiny in February 2012. Of the remaining datasets, DVLA data also showed strong potential for matching a high proportion of the electorate and when combined with DWP data in one area increased the match rate by almost 10%. As the initial purpose of the pilots was not to test the potential of pre-verification limited testing was conducted and further work in a greater number of areas is necessary to explore this further.

Some preliminary analysis which looked at DWP match rates by wards indicated that rates varied by area (as may be expected), with wards with more settled populations having higher match rates than those with transient populations. This indicates that pre-verification will be useful for the majority of electors but further testing is required to establish who is missing from the matches to allow EROs to identify those areas and individuals they should be focusing their efforts on. In addition there is the potential for some dual inaccuracies within the datasets so further testing could look at accuracy in more detail and assess the extent of any potential problems.

Recommendation 8: Further testing is required on data matching for pre-verification, this should include the potential of other datasets to increase the DWP match rate, testing in a variety of area types to allow differences to be explored, and work to assess the accuracy of the data and match rates.

The 22 data matching pilots that took place in 2011 have enabled the Government, DHOs and EROs to learn many invaluable lessons about the process and delivery of matching the electoral register to national datasets. The potential impact of data matching in identifying missing electors has not been proven by this evidence but some datasets have shown potential for identifying specific target groups and should be explored further. The potential for pre-verification offers the chance to ensure a smooth transition to individual electoral registration by keeping the register as complete as possible and further testing is important to explore this. Data matching could prove to be a cost effective mechanism if pre-verification allows the resources of EROs to be freed and used to target the traditionally under-registered and missing electors. Ultimately data matching for the purpose of identifying missing electors may be used by different local areas in different ways, at different times for different purposes – there is unlikely to be a single solution for all areas. Whereas pre-verification has the potential to be of benefit to all local areas during the change to the system of electoral registration.

REFERENCES

Electoral Commission (2005): *Understanding electoral registration. The extent and nature of non-registration in Britain*. August 2005. Electoral Commission: London

Electoral Commission (2010): *The Completeness and accuracy of electoral registers in Great Britain*. March 2010. Electoral Commission: London

Electoral Commission (2011) *Great Britain's Electoral Registers 2011* Electoral Commission London

Fisher, S., Heath, A., Rosenblatt, G., Sanders, D. and Sobolewska, M. (2011): *Ethnic Minority British Election Study: Electoral Registration and turnout data*. October 2011. http://www.runnymedetrust.org/news/368/272/New-data-on-BME-voting-patterns

Cabinet Office Electoral Registration Transformation Programme

Data Matching Schemes

1. Introduction

In this prospectus, the Government invites expressions of interest from English, Welsh and Scottish local authorities who would like to run data matching schemes pursuant to the provisions of sections 35 and 36 of the Political Parties and Elections Act 2009. This prospectus sets out the aims of the intended schemes, what is required of those local authorities who would like to be considered and the subsequent evaluation by the Electoral Commission.

The purpose of these schemes is to gather evidence on whether access to additional data held by public authorities will be useful in helping Electoral Registration Officers (EROs) to maintain and improve electoral registration rates. The schemes may also support EROs by targeting currently under-represented groups and identifying people who are eligible to be registered but are not currently on the register or for whom the details on the electoral register are inaccurate.

The schemes are intended to support the wider work on ensuring and improving the comprehensiveness and accuracy of the electoral register as part of the overall transition to individual electoral registration, and will also identify whether and how access to public authority databases might assist EROs in meeting their duty under section 9 of the Electoral Administration Act 2006.

The development and approval of the schemes will be managed through a Project Board chaired by the Cabinet Office - and including representatives from the Association of Electoral Administrators (AEA) and the Electoral Commission - as part of the wider Electoral Registration Transformation Programme.

We would like to encourage a range of local authorities to run a scheme and hope to see a variety of schemes testing different data sources in a range of contexts. A number of data sets that we hope to have available have already been considered with the help of electoral administrators. They are listed in annex A. The list is indicative of what we hope will be available but we would like to hear from any EROs who have a particular interest in testing any other data set that they think would be useful. Proposals for schemes that address the causes or characteristics of under-registration in specific groups that have a history of under registration will be of particular interest.

Participating in a data matching scheme is likely to benefit you in a range of ways. It may help improve the accuracy and completeness of your electoral register and it will certainly enable you to test whether access to a particular database in your area will help improve your registration rates. Depending on the database in question it may help you to target groups in your area which are consistently under-represented in your register. These might include specific socio-demographic groups, ethnic minority communities or certain age groups.

By 2013 the Government will want to be able to understand:

1. The usefulness of data matching to the longer term sustainability of IER, whether data matching should become part of 'business as usual' and integral to individual electoral registration.

2. The usefulness of data matching in identifying and engaging specific under registered groups, including attainers, service voters, the mobile population, some Black, Asian and Minority Ethnic (BAME) groups and students and

3. How far data matching is effective in capturing changes to the register, in year, including house movers.

If, having read this brief prospectus, you would like to register your interest with Cabinet Office in running a data matching scheme or finding out more please get in touch with XXXX. We need to agree our partners for the data matching schemes by 5 November 2010, so the sooner you are able to get in touch, the better.

e-mail:
telephone:

2. Background

The Political Parties and Elections Act 2009 (the PPE Act) puts in place a statutory timetable for the introduction of individual electoral registration in Great Britain. The Coalition Agreement promises to speed up implementation of Individual Electoral Registration (IER) to tackle electoral fraud. To that end the Government intends to drop the previous Government's plans for a voluntary phase leading up to IER and instead will legislate to bring forward implementation of compulsory IER to 2014, ahead of the next election. To give those already registered at least 12 months to comply with the new requirements of IER no person who fails to register under IER will be removed from the electoral register until after the General Election in May 2015. Any new registrations or changes after implementation in 2014 would need to be carried out under IER. It will also be a requirement from 2014 that anyone wishing to cast a postal or proxy vote should be registered under the IER provisions. Alongside this development, the PPE Act (under sections 35[26] and 36[27]) allows for the establishment of a series of data matching schemes to investigate how increased access to data held by public authorities (which they cannot already access) can assist EROs to maintain and improve the accuracy and completeness of the electoral register; to support the transition to IER by minimising any negative impact it may have on registration rates; to complement the data available to EROs and collected by the annual canvass; and to assist EROs with in year scrutiny of their registers.

The schemes will be expected to test the ability of data matching to identify individuals not registered, with added emphasis on those groups in society who are typically under registered. So, for example, EROs may find that a test of their electoral register against the Department for Work and Pension (DWP) database will enable the identification of a group qualified and present on this database but under registered on the electoral register or who appear on the electoral register but not elsewhere. The schemes will be formally evaluated by the Electoral Commission, and will inform the development of structures to support the shift to full individual registration.

3. Evaluation and Approval of Applications

As part of the application process an ERO will need to identify the public authority or authorities that they wish to obtain data from, and:

- state what data set(s) held by the public authority or authorities the applicant wishes to access;
- detail how the data set(s) might help the ERO meet their registration duty;
- specify how frequently they would require the data set and in what format;
- set out when and for how long the scheme should run (initially in a period between June and early September 2011) and why;
- provide an estimate of the cost, broken down by activity to be undertaken, for example, investment in resources, infrastructure and IT;

[26] Political Parties and Elections Act 2009 – Data Schemes http://www.opsi.gov.uk/acts/acts2009/ukpga_20090012_en_7#pt4-pb3-l1g35
[27] Political Parties and Elections Act 2009 – Schemes under section 35: proposals, consultation and evaluation http://www.opsi.gov.uk/acts/acts2009/ukpga_20090012_en_7#pt4-pb3-l1g36

- set out their objectives in undertaking the scheme;
- set out how they will project manage the scheme
- consider how progress of the scheme will be monitored and risks managed internally and with the Cabinet Office;
- consider how information and data to enable the evaluation of the scheme will be collected and reported;
- set out how to ensure data and IT security; and
- produce or work with Cabinet Office to produce a privacy impact assessment (guidance is available at http://www.justice.gov.uk/guidance/dataprotection.htm.)

Whilst we are aware that EROs working in two tier local authorities are unable to access data held by the upper tier which covers their geographical area, it is not proposed that these schemes will cover access to such data. The purpose of the schemes is to test data sets external to the local authority that EROs currently cannot access to inform our understanding of their usefulness in enhancing the accuracy and comprehensiveness of the register. We are aware that some EROs in two tier local authorities would like parity of access with those in unitary authorities and that is a reform that is being considered outside of the scope of this project.

During the planning for data matching schemes we have identified a range of potential data sources and they are listed at annex A. We welcome ideas from EROs about other data sets that would be useful, either local or national. EROs are particularly encouraged to consider the potential of any other local sources of data for a data matching scheme.

Cabinet Office is already discussing the use of the data sets with the public authorities that hold them. Schemes will only proceed after consultation on the data sets with the data holder, the Electoral Commission and the Information Commissioner has been concluded. This may amount to a significant part of the development time of a scheme and should be taken in to account when electoral administrators are developing a proposal.

In addition to evaluating proposals to run schemes on the basis of how they address the criteria above, the Cabinet Office will need to be satisfied that:

- effective project management arrangements will be in place and sufficient resources and support within the local authority and any external suppliers will be available to support and deliver the scheme;
- there is local political support for the scheme;
- the formal evaluation process undertaken by the Electoral Commission will be fully supported;
- the costings for the scheme are on a value for money basis; and
- the scheme will provide us with evidence to inform the work on enhancing the comprehensiveness and accuracy of the register and targeting of under-represented groups.

It will be particularly important for proposals to take account of the need to maintain public confidence in the use of data held by public authorities and the electoral process.

In considering approval of the overall package of data matching schemes, the Cabinet Office may also take into account the following in order to get an effective range of schemes and maximise learning from them:

- the geographic spread of authorities involved in data matching schemes;
- the types of authorities (for example, metropolitan, rural and unitary); and
- the range of public authority databases involved.

We would like to hear from EROs interested in testing the following:

Data held by Driver and Vehicle Licence Agency (DVLA) (provisional licence holders), Her Majesty's Revenue and Customs (HMRC) (child benefit), the National Pupil Database. EROs will be asked to consider, in conjunction with the data holders and with Cabinet Office, how these data sets could be used in testing schemes to identify or validate attainers already registered, attainers who should not be on the register (non eligible) or who should be registered and are not.

The usefulness of HMRC's national insurance and PAYE recording system and DWP's customer information system (CISx). These databases hold the name, address and date of birth of those employed or in receipt of a pension, or on a lower earnings limit. They include a high proportion of young people, BAME and the mobile population. EROs will be asked to consider, in conjunction with the data holders and with the Cabinet Office, how these data sets could be used to identify invalid duplications, inaccurate entries and those who are present on databases other than the electoral register who are eligible to register.

The Royal Mail national change of address update contains names and addresses (old and new) of individuals who have moved or are in process of moving house. The database is collected directly from the Royal Mail redirection application forms completed by customers all over the UK in the process of moving home. EROs should consider using this data to identify home movers.

The Ministry Of Defence (MOD) joint personnel administration system (JPA) holds national insurance number, name, date of birth and address of members of the Armed Forces. EROs should consider using this data to identify service personnel and duplications.
Cabinet Office will provide support and guidance to EROs who intend to register their interest in participating in a data matching scheme – please do not hesitate to contact us if you wish to discuss the above criteria and the application process.

4. Preparation of data matching scheme orders

Assuming the above criteria are met, data matching schemes will be enabled by legislative orders and most likely supported by memoranda of understanding formalising arrangements between the local authority and the data holder. Any order will set out the arrangements around access, control and use of the data sets and will be subject to the affirmative resolution procedure which requires Parliament to debate the order in both the House of Commons and the House of Lords.

It may be that there is scope for more than one authority seeking access to the same data set to be included in one order or for more than one data set to be accessed by one or more authorities.

An order will give legislative authorisation for the data sharing but it is anticipated that those sharing data under any scheme made by order will have regard to the effect of Article 8 of the ECHR, the common law of confidence and any relevant provisions of the Data Protection Act 1998.

Section 36 of the PPE Act creates a number of procedural steps which must be followed before an order under section 35 can be made to create a scheme. It provides that a scheme can only be created where an electoral registration officer has submitted a proposal to the Deputy Prime Minister for consideration and the Deputy Prime Minister approves that proposal or does so with modifications agreed to by the registration officer.

Before making an order, the Deputy Prime Minister must consult (a) the Electoral Commission, (b) the body that is authorised or required by the order to provide data to the ERO and (c) the Information Commissioner. There is also a requirement that each order must include a specific evaluation date and that the Electoral Commission must prepare an evaluation report on that scheme.

Cabinet Office will plan and manage the laying of such orders but may require local authority input during their preparation. In particular, each scheme will require a privacy impact assessment (details of which can be found via the link later in this document).

5. Reporting on data matching schemes as part of Cabinet Office's wider Electoral Registration Transformation Programme

Cabinet Office will co-ordinate the high-level project management requirements for the schemes as a whole. Cabinet Office will negotiate service level agreements with the local authorities undertaking schemes and expect regular highlight reports and minutes of local project board meetings.

Each scheme will also be required to report exceptions to the Cabinet Office project manager who may agree corrective action and/or escalate as appropriate within the overall programme structure.

6. Evaluation of the data matching schemes by the Electoral Commission

The Electoral Commission is responsible for the evaluation of each scheme, the requirements of which are set out in section 36 of the PPE Act. The Electoral Commission's report on each scheme will evaluate the extent to which the scheme has enabled the registration officer to meet the registration objectives, set out in section 31 (8) of the PPE Act:

- that persons who are entitled to be registered in a register are registered in it,
- persons who are not entitled to be registered in a register are not registered in it; and
- that none of the information relating to a registered person that appears in a register or other record kept by a registration officer is false.

The evaluation is likely to include the administrative, financial and other demands of each scheme and any objections to the scheme, for example, from members of the public.

The registration officer must give the Electoral Commission such assistance as they may reasonably require while preparing the report and on receipt of the report from the Electoral Commission, the registration officer must publish it as they deem appropriate.

For further information on the evaluation process please contact XXXX at the Electoral Commission.

e-mail
telephone:

7. Funding

Cabinet Office will want to consider the full costs of each scheme in order to ensure proportionality and effective use of public monies before any approval by the Deputy Prime Minister.

The following costs will be covered as part of a data matching scheme:

- Any additional administrative costs to the ERO and the public authority in providing the information in the format required, and the consequent use of it.
- Any additional administrative cost to the ERO for providing data to the Electoral Commission for evaluation
- Any additional necessary licence costs for access to the requested data.
- Costs associated with ensuring that information security standards are maintained by all parties.
- Any further necessary additional costs relating to the schemes, including costs of collecting and providing evaluation data, approved in advance.

If you are in any doubt about funding of any elements of a data matching scheme you are thinking of proposing please talk this through with the team at Cabinet Office (contact details at page 3). Funding for the schemes will be provided following their implementation.

8. Outline timetable

The following provisional timetable indicates the target dates for key milestones:

From issue of the prospectus – ongoing discussion with EROs and public authority data holders about data sources, data handling and data security issues. Work on data security with other governmental agencies and prospective public authorities to be involved in schemes.

27 September onwards — Expressions of interest from EROs made to Cabinet Office

Cabinet Office project team available to electoral services officers for consultation on data matching schemes – discussion, design, planning.

5 November — Closing date for expressions of interest and initial proposals

November onwards — Liaison between Cabinet Office and EROs to agree and prepare for schemes

Preparation and laying of Statutory Instrument(s) enabling the schemes

June 2011 onwards — Data matching schemes in operation

Annex A [of prospectus]

Potential data sources

EROs can, and will, only have access to names, addresses date of birth/age (where appropriate) and nationality.

EROs already access some or all of the following data sets:

- The register of deaths
- Council register records
- Registers of households in multiple occupations
- Local land and property gazetteers
- Housing benefit applications
- List of persons in residential and care homes; and where allowed
- Details of attainers (those aged 16 or 17) held by educational departments.

Additional data sets that have been suggested by electoral administrators that could be useful in identifying people not already on the register or could provide more accurate or up to date information:

- DVLA provisional licence database
- Child benefit database
- National insurance and PAYE recording system
- DWP customer information system and housing benefit
- National pupil database (NPD)
- Royal Mail national change address update
- MOD joint personnel administration system

Annex B [of prospectus]

Additional information

You can find further information on individual electoral registration that may be useful in preparing a proposal in the following documents and links:

Cabinet Office web pages

http://www.cabinetoffice.gov.uk

http://www.justice.gov.uk/news/newsrelease200709b.htm

Political Parties and Elections Act:

http://www.justice.gov.uk/publications/political-parties-elections-bill.htm

http://www.opsi.gov.uk/acts/acts2009/ukpga_20090012_en_1

Data Protection including privacy impact assessments

http://www.justice.gov.uk/guidance/dataprotection.htm

Hansard on PPE Bill

http://services.parliament.uk/bills/200809/politicalpartiesandelections/stages.html

Information Commissioner –

http://www.ico.gov.uk/

EC web pages on evaluation and IER –

http://www.electoralcommission.org.uk

AEA web pages

Rural Urban classifications for English Local Authorities

	Pilot No's	Overall No's	Pilot %	Overall %
Large Urban	1	39	5%	12%
Major Urban	9	71	47%	22%
Other Urban	2	58	11%	18%
Rural 50-80%	2	48	11%	15%
Rural 80%+	4	55	21%	17%
Significant Rural	1	55	5%	17%

Source: Rural/Urban Local Authority (LA) Classification (England), Office for National Statistics
http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/rural-urban-definition-and-la/rural-urban-local-authority--la--classification--england-/index.html

Percentage of Scottish population in each Urban Rural Classification - Council Areas included in the data matching pilot compared to Scotland total

	Large Urban	Other Urban	Access-ible Small Towns	Remote Small Towns	Access-ible Rural	Remote Rural
Renfrewshire Joint Valuation Board						
Renfrewshire	76.0	10.1	9.5	0.0	4.4	0.0
East Renfrewshire	86.5	0.0	9.5	0.0	4.0	0.0
Inverclyde	0.0	86.4	8.0	0.0	5.6	0.0
Lothian Joint Valuation Board						
Edinburgh, City of	96.3	0.0	2.6	0.0	1.1	0.0
East Lothian	23.3	10.8	23.3	15.0	24.7	2.9
West Lothian	0.0	81.3	8.9	0.0	9.8	0.0
Midlothian	0.0	68.2	14.5	0.0	17.3	0.0
Glasgow City	99.8	0.0	0.0	0.0	0.2	0.0
Total Scotland	**38.9**	**30.6**	**8.5**	**3.8**	**11.6**	**6.5**

Source: Urban Rural Classification 2009-2010 Population Tables, General Register Office for Scotland
http://www.scotland.gov.uk/Topics/Statistics/About/Methodology/URtables2010

Interview topic guide for pilot sites – EROs/Pilot lead

Introduction
My name is [ ] and I'm a researcher working for the Cabinet Office. I'm part of a team who are currently undertaking an evaluation of the data matching pilots you have been taking part in. All the information that you provide will be treated as confidential and when using quotations no individual or organisation names will appear in any reports, so please be open and frank about your views. We will not pass on what you say to anyone in your organisation.

The interview should last for a maximum of one hour and will include a number of questions for your consideration. I would like to record the interview as this is more accurate than taking notes, is that ok?

(*If not, ask what their concerns are and try to reassure them. If need to tell them that the tapes will be transcribed and destroyed at the end of the study; tapes will be held securely in the CO; CO is a professional body; interviews should be covered by the Data Protection Act*).

Are there any questions you would like to ask before we start?

Setting the scene
As I have said I will be asking questions about the data matching pilots and your experience of the pilot as well as your views on their potential impact on the electoral register and general usefulness. But I would just like to ask you a few questions about your role and local authority to begin with...

1. What is your current role?
   - *Main responsibilities and how this relates to electoral registration?*

2. How long have you been in your current role?

3. What has been your role in the data matching pilot?

4. How many people work on electoral registration in your local authority?
   - *What are their roles and responsibilities?*
   - *Do you think this resource is currently adequate?*

Setting up the pilot
Ok I'm now going to ask you some questions about the process of setting up the pilot...

5. Why did you or your local authority decide to take part in the data matching pilots?
   - *Particular interest or problem in the area? E.g. fraud*
   - *Keen on IT?*
   - *Want to improve overall registration rates or target particular groups?*

6. What did you hope to achieve from the data matching pilots?
   - *Increased completeness or accuracy?*
   - *Increased in under-registered groups – if so which?*
   - *Target fraud?*

7. How did you decide on and develop the approach in your original proposal?
   - *Ever done local data matching before?*
   - *Particular interest in groups?*
   - *Resources available to complete work?*

8. How did you find the process of setting up the data matching pilot?
   - *Legalities? E.g. Art 4 and PIA*

- *Secure emails?*
- *Methodology?*
- *Guidance and communication from CO?*

9. How much time was allowed for this process and was this long enough from your perspective?
   - *What were main delays if any? And why?*

10. Before the pilot began, how well informed did you feel about what data you could expect to receive and what you would then do with this data?
    - *What other information would like to have had at this stage?*
    - *More support needed in terms of methodology and evaluation?*

Pilot initial stages – matching and analysing the data
I'd now like to move on to asking you some questions about the first stages of the pilot; receiving and analysing the matched data....

11. How would you describe your experience of sending and receiving data for your pilot? If you matched to different datasets please try to tell me about any differences in each of these experiences.
    - *IT issues – within EMS or by DHO or both, what – secure emails, firewalls, file sizes, software issues etc.*
    - *Lack of clarity or confusion around the process?*

12. When you received the data how easy was it to understand and use?
    - *As above try to separate out views of different data sets where appropriate.*
    - *Why – more guidance needed to explain data?*
    - *File format too complicated?*
    - *Scale of data?*
    - *Technical skills or resources available in authority?*

13. What did you think of the quality of the data and the match scores?
    - *Currency?*
    - *Accuracy of match scores – too high, too low, false positives or negatives etc?*

14. What, if anything, would you have changed about the data you received?
    - *Updates from a shorter time period?*
    - *Mismatches only?*
    - *More sophisticated/accurate matching coding.*
    - *Format of data?*
    - *Impact of time delays?*
    - *Timing e.g. clashed with canvass?*

15. How long did it take and what resources did you need to clean and analyse the data?
    - *Split this out for different datasets if appropriate and if a difference*
    - *Assistance from EMS supplier?*
    - *Other technical support guidance?*
    - *Have good technical skills.*
    - *Resources in place or had to 'buy in'?*
    - *Ways in which could have taken less time?*

16. What, if any, local matching did you carry out?
    - *What was the purpose of this?*
    - *How long did it take?*
    - *Any problems or issues? I.e. compatibility, technical skills etc.*

Follow up work
I'm now going to move on to discuss the work you did to follow up the data and how this related to your canvass....

17. Once you had completed and analysed your data, how did you approach your follow up work?
    - *How did they decide which people to follow up and how e.g. random %, people with certain match scores etc, wrote to them or door knocked or both?*
    - *Resources available? Scale of data?*
    - *How did this relate to your canvass e.g. pre, post, during, reminder stage etc.*
    - *Did you have a control group? What size and how selected?*
    - *How, if at all did this vary to your original plan and why?*

18. At this stage of your pilot, how effective do you feel each of the data sources you have matched to have been in identifying people who were not on the electoral register but were entitled to be?
    - *Views on how this compares to canvass activity?*
    - *Views on time, effort, cost?*
    - *Views on comparability to local data sources?*

19. At this stage of your pilot, how effective do you feel each of the data sources you have matched to have been in identifying people who should not have been on the electoral register?
    - *Views on how this compares to canvass activity?*
    - *Views on time, effort, cost?*
    - *Views on comparability to local data sources?*

20. At this stage of your pilot, how effective do you feel each of the data sources you have matched to have been in helping to assess the accuracy of your electoral register?
    - *Views on how this compares to canvass activity?*
    - *Views on time, effort, cost?*
    - *Views on comparability to local data sources?*

21. At this stage of your pilot, how effective do you feel each of the data sources you have matched to have been in helping to reach your target group(s)?
    - *Views on how this compares to canvass activity?*
    - *Views on time, effort, cost?*
    - *Views on comparability to local data sources?*
    - *Under-registered groups in general?*

22. What, if any, feedback have you received from members of the public about the pilot?
    - *Positive, negative?*
    - *Who and how? I.e. people who they're written to specifically or people who've read about it in local press etc.?*
    - *Phone calls, questionnaires etc?*

Future of data matching
I'd like to finish by asking some questions about the possible future of data matching....

23. How, if at all, would you like to see data matching work in the future?
    - *Time of year.*
    - *Currency of data*
    - *Mismatches only?*
    - *Use of verification or identifying missing electors or both?*
    - *Data sources – in particular local and national?*

24. Is there anything else you don't think I have covered or would like to add?

Thank you very much for your time.

Publication date: March 2012

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