



ONS Census Transformation Programme

Annual assessment of
ONS's progress towards an
Administrative Data Census
post-2021

June 2017

Table of contents

| | |
|---|----|
| 1. Main points | 2 |
| 2. Why are we doing this? | 5 |
| 3. Things you need to know about the assessment | 6 |
| 4. What's changed since last year's assessment | 7 |
| 5. What needs to be in place to move to an Administrative Data Census? | 9 |
| 6. Current assessment of ONS's ability to move to an Administrative Data Census | 11 |
| 6.1 Access to data | 12 |
| 6.2 Ability to link | 15 |
| 6.3 Ability to meet information needs of users | 17 |
| 6.3.1 Population estimates | 17 |
| 6.3.2 Households and families | 19 |
| 6.3.3 Population, housing and household characteristics | 21 |
| 6.4 Acceptability to stakeholders | 26 |
| 6.5 Value for money | 29 |
| 7. Conclusions and next steps | 30 |
| | |
| Annex A. High level evaluation criteria explained | 31 |
| Annex B. Update on acquisition of data | 33 |
| Annex C. A Quality Framework for Admin-based characteristics | 39 |

1. Main points

This year's assessment reflects an important step forward in the Office for National Statistics (ONS) being able to access the range of data needed to produce Administrative Data Census outputs. The Digital Economy Act 2017 was passed into law in April 2017. The Act gives ONS a right of access to information held by government departments, other public bodies, charities and large and medium-sized businesses, for statistics and research purposes.

Over the last year, we've done a lot of ground work to enable us to achieve our expected assessment post 2021 (see Figure 1). As well as the Digital Economy Act 2017, progress has been made in:

- improving the accuracy of administrative data based population estimates through improved linking, use of new data sources and improved methodology; these population estimates are now closer to our official estimates
- producing new estimates on the number of occupied addresses ("households") and estimates of personal income direct from administrative data sources
- identifying potential administrative sources and starting to describe methods which will enable us to produce estimates of population and household characteristics using these sources in combination with survey and other data

Our plans for the next year build on this ground work. They continue to demonstrate the potential for an Administrative Data Census to produce outputs that meet information needs. This will involve working with data suppliers to access data sources through the new powers of the Digital Economy Act 2017. The Act will enable us to make progress against the other high-level criteria over the coming years. This progress is reflected in the "expected progress by 2018" indicator in the main assessment as shown in Figure 1.

Figure 1. Overall assessment against five high level criteria

Key to chart

→ Some progress

↗ Good progress

△ Change in assessment

| Evaluation criteria | | 2016 assessment | Progress made since 2016 | 2017 assessment | Expected progress by 2018 | Expected assessment by 2023 |
|---|--|-----------------|--------------------------|-----------------|---------------------------|-----------------------------|
| Access to data | | RED/ AMBER | △ | AMBER | ↗ | AMBER/ GREEN |
| Ability to link | | AMBER | ↗ | AMBER | ↗ | GREEN |
| Ability to meet information needs of users: | Population estimates | AMBER | ↗ | AMBER | ↗ | GREEN |
| | Households and families estimates | RED/ AMBER | ↗ | RED/ AMBER | ↗ | AMBER/ GREEN |
| | Population and household characteristics | RED/ AMBER | → | RED/ AMBER | → | AMBER |
| | Housing characteristics | AMBER | → | AMBER | → | GREEN |
| Acceptability to stakeholders | | RED/ AMBER | → | RED/ AMBER | → | AMBER/ GREEN |
| Value for money | | AMBER | → | AMBER | → | AMBER/ GREEN |

Notes on table: The expected progress by 2023 has not changed since last year

In particular, we'll be working within the legally operational framework of the Digital Economy Act 2017 to:

- access more 'activity'¹ data, which will be used in combination with a coverage survey to improve the quality of population estimates
- access a wider range of data covering population characteristics and use it to increasingly demonstrate the range of outputs that are possible through an Administrative Data Census

Together with accessing data sources, we will develop new methods to enable the production of small area multivariate² outputs about population characteristics.

¹ Information from administrative data sources about when individuals have interacted with systems or services, such as the National Insurance, tax or benefit systems, or a hospital visit through the NHS system.

² Cross tabulated outputs using more than one variable, for example unemployment rates by ethnic group for small areas.

2. Why are we doing this?

In March 2014, the National Statistician made a recommendation that the census in 2021 should be predominantly online, making increased use of administrative data and surveys to both enhance the statistics from the 2021 Census and improve statistics between censuses. The government's response to this recommendation was an ambition that "censuses after 2021 will be conducted using other sources of data".

A move towards an Administrative Data Census is our response to this challenge.

An Administrative Data Census offers a number of opportunities. These include producing key census outputs on a more timely, frequent basis (possibly every year), for less money than the current system. It also offers opportunities to produce new census outputs that aren't available through the current approach, such as income, fuel poverty and housing affordability. This will result in better decision-making across government in line with UK Statistics Authority strategy – [Better Statistics, Better Decisions](#).

3. Things you need to know about the assessment

This is ONS's second assessment of its ability to move to an Administrative Data Census in the next decade.

It is our ambition to produce the type of information that is collected by a 10-yearly census (on housing, households and people) from an Administrative Data Census. Doing this will require a combination of:

- record-level administrative data held by government
- a population coverage survey
- a population characteristics survey
- some commercial and other non-survey data sources

You can find further information on what an Administrative Data Census is in [last year's assessment](#).

The assessment in this report is made against the following five high-level criteria:

1. Rapid access to new and existing data sources
2. The ability to link data efficiently and accurately
3. Methods to produce statistical outputs that meet priority information needs of users
4. Acceptability to stakeholders
5. Value for money

These criteria reflect what needs to be in place for ONS to move to an Administrative Data Census. Further details on these criteria can be found in **Annex A**.

4. What's changed since last year's assessment

In the previous report, the focus of our assessment was whether ONS would be able to move to an Administrative Data Census post-2021. We described the goal of comparing outputs based on administrative data and targeted surveys against the 2021 Census.

We have now developed our plans for this comparison in more detail.

To make this comparison as fair and robust as possible, we will need to produce the best possible Administrative Data Census outputs in 2021. We plan to have the following in place by 2021:

Administrative Data Census-based population statistics by 2020

An Administrative Data Census-based approach to producing population statistics using a combination of administrative data and a population coverage survey will need to be in place in 2020 to demonstrate the ability to produce:

- annual estimates of the size of the usual resident population by age and sex for national, local and small areas
- components of population change (births, deaths and migration)

These estimates provide a base for further outputs about the population including population projections.

The annual Population Coverage Survey (PCS) would measure and adjust for coverage errors in the Statistical Population Dataset (SPD) as shown in the framework in Figure 2. We are developing plans to test a PCS, to commence later in 2017. In the next 3 years, we plan to implement a comprehensive testing strategy to explore appropriate sampling and data collection methods, response rates and estimation methodologies.

Earlier this year, we produced a first set of research outputs on the [numbers of occupied addresses \("households"\)](#). These were produced from the same resident population base as that used for Administrative Data Census Outputs on the size of the population.

An Administrative Data Census producing characteristics of the population, housing and households by 2021, supported by an integrated approach to collecting survey information on characteristics of the population

This approach will need methods that make best use of a combination of administrative, survey and commercial data to produce outputs about household and population characteristics. A range of methods may be needed, depending on the availability and quality of these sources. Some topics will be predominantly administrative data-based, others will be based on survey information alone. We expect that for most characteristics we will need to integrate both administrative and survey sources, with support from commercial data in some cases.

For census topics for which there is limited or no administrative data available, such as the number of hours of unpaid care, outputs would be largely based on a sample survey. Outputs for these topics may be more limited in frequency and/or detail. For example for such topics, it might be possible to produce estimates only at Local Authority level. We are currently investigating how such a survey might fit alongside other ONS surveys.

The systems, services and technologies in place to support this transformation.

We are exploring how best to make use of new corporate platforms for the acquisition, management, integration, processing, and analysis of data from multiple sources.

We're in the early stages of development for a range of new technologies to support the transition to an Administrative Data Census including:

- online collection of census and survey data
- acquisition and validation of administrative data directly from suppliers
- matching services for an address and business "spine" and for data about people
- access control and data management

Future assessments will measure our progress towards achieving these objectives. The final assessment (due in spring 2023) will form the basis for the National Statistician's recommendation at the end of 2023, about the future of the census and population statistics. A public consultation will take place on the basis of our final assessment and to ensure the National Statistician's recommendation reflects user needs.

5. What needs to be in place to move to an Administrative Data Census?

Figure 2 shows a framework for operating an Administrative Data Census. It sets out how each of the components fit together, as follows:

- a range of administrative, survey and other data sources are combined using linkage methods (our current methods are described in our [methodology paper](#))
- we then apply a [set of rules and methods](#) to create a Statistical Population Dataset (SPD)
- we will then link the SPD to a Population Coverage Survey and use estimation methods to produce outputs on the size of the population and households
- we will also link to a characteristics survey, and use other methods to produce outputs about the characteristics of the population

Figure 2 has been mapped to the high level evaluation criteria (see **Annex A**) and coloured using the Red Amber Green (RAG) status to illustrate our progress. It shows that having access to a few key data sources enables us to produce population and household estimates for the total population (denominators) that are fairly close to the official estimates. We have published two sets of research outputs on population estimates in October 2015 and November 2016 – around 95% of administrative data based population estimates for local authorities are of similar quality to those produced by the Census in 2011.

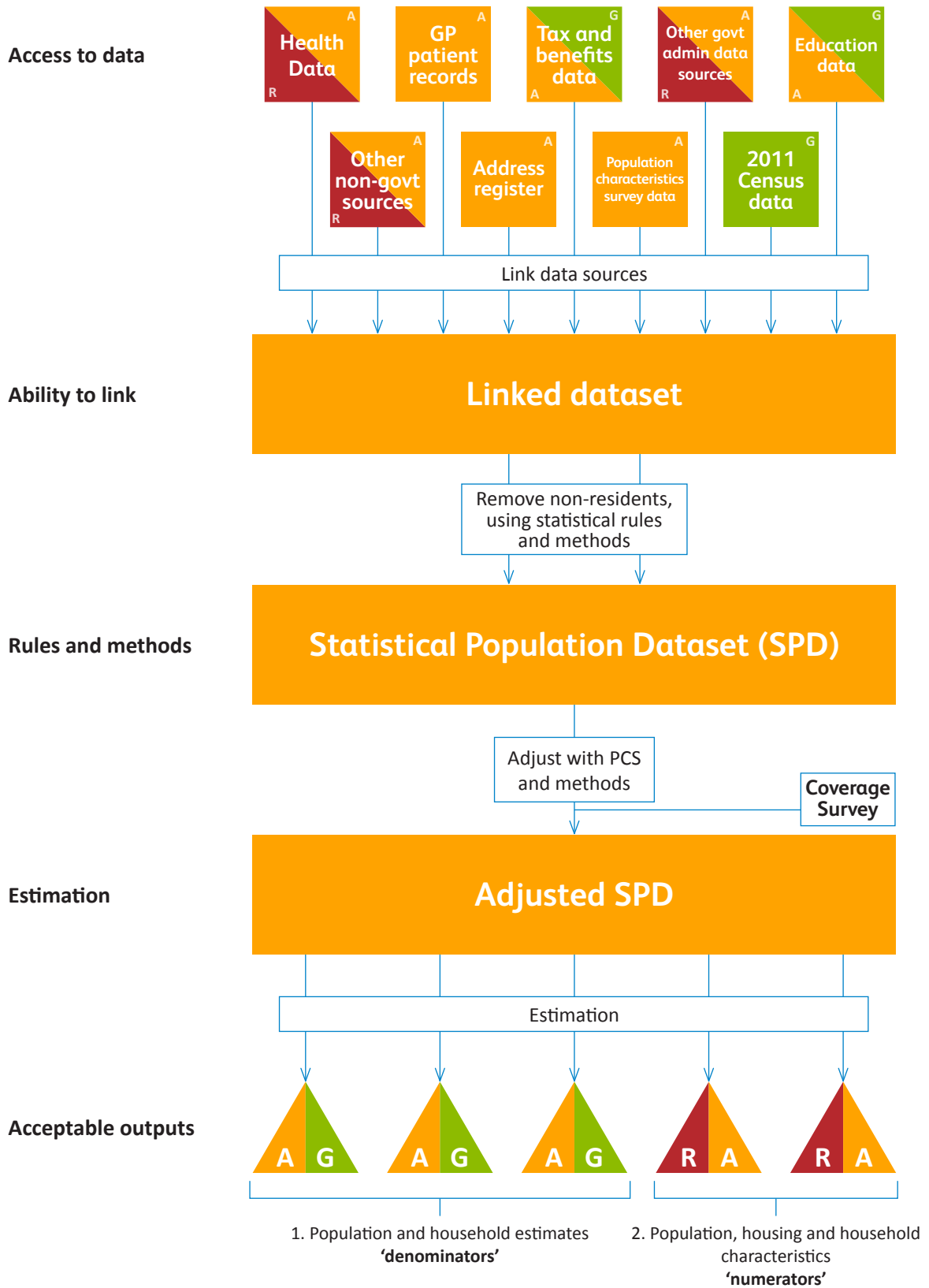
We also published a first set of estimates of occupied addresses (“households”) from administrative data in February 2017, again showing good potential. This is reflected in the diagram by the green/amber colour.

Our greatest challenge is our ability to produce estimates of the characteristics of the population and households (numerators), currently red/amber on the diagram.

Our current ability to produce statistics on the topics traditionally collected in the census is directly related to the limited access that we currently have to administrative sources that have information about such characteristics (coloured red/amber). Gaining access to more key data sources, using the new powers in the Digital Economy Act 2017, together with developing new methods to produce such outputs will improve our progress in this area.

The next section provides a more detailed assessment on our progress to date and expected progress over the next year against each of the high-level criteria.

Figure 2. Administrative Data Census Framework shaded by RAG status



6. Current assessment of ONS's ability to move to an Administrative Data Census

This high level assessment allows a direct comparison to be made with last year's assessment. Work is ongoing to produce more detailed criteria. This includes describing definitions for the Red Amber Green (RAG) status for each of the criteria Figure 3 shows the assessment demonstrating where we are now and where we expect to be by 2023. The rest of this section provides the evidence behind the assessment and a description of what will be done in the future to improve the assessment. Each assessment indicates the expected progress over the next year.



Figure 3. Current assessment of ONS's ability to move to an Administrative Data Census

Key to chart

→ Some progress ↗ Good progress △ Change in assessment

| Evaluation criteria | 2016 assessment | Progress made since 2016 | 2017 assessment | Expected progress by 2018 | Expected assessment by 2023 | |
|---|--|--------------------------|-----------------|---------------------------|-----------------------------|--------------|
| Access to data | RED/ AMBER | △ | AMBER | ↗ | AMBER/ GREEN | |
| Ability to link | AMBER | ↗ | AMBER | ↗ | GREEN | |
| Ability to meet information needs of users: | Population estimates | ↗ | AMBER | ↗ | GREEN | |
| | Households and families estimates | RED/ AMBER | ↗ | RED/ AMBER | ↗ | AMBER/ GREEN |
| | Population and household characteristics | RED/ AMBER | → | RED/ AMBER | → | AMBER |
| | Housing characteristics | AMBER | → | AMBER | → | GREEN |
| | Acceptability to stakeholders | RED/ AMBER | → | RED/ AMBER | → | AMBER/ GREEN |
| Value for money | AMBER | → | AMBER | → | AMBER/ GREEN | |

6.1 Access to data

| Evaluation criteria | 2016 assessment | Progress made since 2016 | 2017 assessment | Expected progress by 2018 | Expected assessment by 2023 |
|---------------------|-----------------|---|-----------------|---|-----------------------------|
| Access to data | RED/AMBER |  | AMBER |  | AMBER/GREEN |

Evidence

The [Digital Economy Bill](#) received Royal Assent in April 2017. The Digital Economy Act 2017 gives the UK Statistics Authority a statutory right of access to information held by government departments, other public bodies, charities, and large and medium-sized businesses, for statistics and research purposes. This will help ensure ONS has access to the data it needs to produce fit-for-purpose official statistics that meet the challenges of a modern administration and the evolving needs of statistical users.

Figure 4 shows the current availability of administrative and other non-survey data sources for the key topics traditionally included in the census. This is based on recent exploratory work to assess the potential for administrative and other sources to provide information about characteristics of the population.

In this table, a green assessment means that there are some data currently available to us on those topics. This doesn't necessarily mean that there is enough data to produce direct outputs based on administrative data alone. In many cases, these data sources would need to be combined with survey sources to produce characteristic outputs.

An amber assessment has been given to those topics where administrative data may be available, but ONS does not currently have access to it. A red assessment indicates topics that we don't think are covered on any administrative data sources. In particular, this highlights the challenges for certain topics such as "mode of travel to work" and "hours of unpaid care", where there are no data sources available.

The table demonstrates that there are potentially data sources that collect useful information for most topics. However, more work is needed to determine whether the identified sources are of a suitable quality to produce outputs that are fit for purpose. A preliminary assessment of the quality of these sources is presented in the Population, housing and household characteristics section.

Figure 4. Availability of administrative data sources by census topic

This table represents a view of the data available to ONS. It does not capture an assessment of quality, such as coverage, accuracy and relevance.

| GREEN | AMBER | RED |
|---|--|---|
| <p style="text-align: center;">Some data available to ONS</p> <p>Demographics Age Sex Marital or legal partnership status</p> <p>Education Term time address</p> <p>Ethnicity, Identity, Language & Religion Ethnic group Citizenship (passport held) or nationality</p> <p>Health Disability and long term health conditions</p> <p>Housing Accommodation type Number of rooms Number of bedrooms Tenure and landlord (if renting)</p> <p>Labour Market Employed (including students, excluding self-employed) Economically inactive, unable to work i.e. long term sick Unemployed (including students)</p> <p>Migration Country of birth</p> | <p style="text-align: center;">Some data available but ONS currently doesn't have access</p> <p>Demographics Household composition</p> <p>Education Qualifications held</p> <p>Ethnicity, Identity, Language & Religion Ethnic group Citizenship (passport held) or nationality English language proficiency Welsh language Main languages used</p> <p>Health Disability and long term health conditions</p> <p>Housing Tenure and landlord (if renting) Second residences</p> <p>Labour Market Number of hours worked Industry Year last worked Economically inactive, retired</p> <p>Migration Country of birth Internal/international migration (including address one year ago)</p> <p>Travel Number of cars/vans</p> | <p style="text-align: center;">Limited or no data available</p> <p>Demographics Family relationships</p> <p>Ethnicity, Identity, Language & Religion National identity Religion</p> <p>Health General health Amount of unpaid care provided</p> <p>Housing Self-containment of accommodation</p> <p>Labour Market Economically inactive, looking after family Address of place of work Supervisor status Occupation</p> <p>Travel Method of transport to place of work</p> |

Note: for some characteristics, there are multiple data sources available. Therefore, they may appear on the table twice.

Since last year's report, we've assessed the statistical quality of two new data sources. We published our findings in [data source overviews](#) (one covers income and benefits data from the Department for Work and Pension (DWP) and HM Revenue and Customs (HMRC) and the other covers the Personal Demographic System (PDS) data from NHS Digital). We used these sources to produce new research outputs on income and improved estimates of the size of the population.

What we're doing to improve the assessment

Since the Digital Economy Act 2017 has passed into law, the UK Statistics Authority and ONS are putting together a new legally operational framework for sharing data. This will include developing new codes of practice and setting out high level principles that will guide the exercise of new powers.

Annex B provides a list of administrative data sources by topic.

Once we start to obtain access to new sources we will be able to carry out further statistical research to improve our current methods and our ability to produce new statistical outputs.

As described earlier, some variables are not fully available on administrative data. For key topics, one solution might be to explore whether it is possible to collect these topics on administrative data, or if they are already collected, ensure this is done on a consistent basis. An example of this could be to improve the collection of data about ethnicity across the health service.

The assessment is now at amber, due to the Digital Economy Act 2017. The expected assessment of amber/green is on the basis that we can now access the required data sources and that they are of the expected and required quality to produce outputs.

6.2 Ability to link

| Evaluation criteria | 2016 assessment | Progress made since 2016 | 2017 assessment | Expected progress by 2018 | Expected assessment by 2023 |
|---------------------|-----------------|--------------------------|-----------------|---------------------------|-----------------------------|
| Ability to link | AMBER | ↗ | AMBER | ↗ | GREEN |

Evidence

The [Independent Review of Methodology](#) gave two recommendations on data linkage:

- “We encourage further development of matching methods...”
- “The requirements for anonymisation need to be reviewed and the possibility of conducting some research in a safe setting under less stringent anonymisation considered.”

These recommendations are being progressed in the following ways.

Improvements to the accuracy of data matching methods have improved the latest [Administrative Data Research Outputs](#). Probabilistic matching has been used to identify further links between the NHS Patient Register and Customer Information Service³ data. The improved linkage methods including the development of a “statistical spine” are described in more detail in our [methodology paper](#).

Administrative Data Census Research Outputs have been produced for the number of occupied addresses. This was possible through improved methods for matching addresses on administrative records to AddressBase⁴ - a list of residential and other types of address. In linking records to AddressBase, we can then assign a unique property reference number (UPRN), which enables better quality matching of addresses between sources. This method is described in more detail in the methodology section of the [“Households” Research Outputs paper](#).

A proposal for a “trusted third party” linking model (as recommended by the Administrative Data Taskforce in a [report](#) published in 2012) has received support in the recent consultation on the [Better Use of Data in Government](#), which closed in April 2016.

The trusted third party model offers an alternative approach to preserving privacy when linking multiple administrative datasets, while still allowing us to improve the quality of matching and meet statistical objectives. In summary, this approach involves the separation of person identifiers (names, dates of birth and addresses) from information about their characteristics. Record linkage can then be undertaken to the highest standards while still adhering to our principle of not holding identifiable information in one place for longer than required.

³ Customer Information System of the Department for Work and Pensions, an administrative data source which contains a list of people who have a National Insurance number.

⁴ AddressBase – an Ordnance Survey address product compiled from local authority, Ordnance Survey and Royal Mail address lists.

[The consultation response report](#) indicates there was strong support for the proposal and for “clear and robust safeguards to promote assurance”.

What we're doing to improve the assessment

We are working with data suppliers to develop common standards on Government-held data, aligning with priority principles set out by the [cross-Government Data Programme](#).

Adding the UPRN⁵ to the administrative sources lets us identify different records for the same address. We are continuing to develop our address matching processes which may be used across government to help improve the quality and consistency of addresses held on administrative data.

⁵ A unique property reference number (UPRN) is a unique alphanumeric identifier for every spatial address in Great Britain and can be found in Ordnance Survey's address products.

6.3 Ability to meet information needs of users

6.3.1 Population estimates

| Evaluation criteria | 2016 assessment | Progress made since 2016 | 2017 assessment | Expected progress by 2018 | Expected assessment by 2023 |
|----------------------|-----------------|--------------------------|-----------------|---------------------------|-----------------------------|
| Population estimates | AMBER | ↗ | AMBER | ↗ | GREEN |

Evidence

The second set of Administrative Data Census Research Outputs on the size of the population was published in November 2016. Estimates were provided down to Lower Layer Super Output Area (LSOA) and by single year of age (at local authority level) for 2011 and 2015, in response to user feedback.

We also produced an improved version of the Statistical Population Dataset (SPD) v2.0 which generally produced estimates that were more similar to the 2011 Census estimates than those produced from SPD v1.0. Some of the most notable improvements have been for the female population and children aged 5 to 14. Improvements are largely as a result of:

- including school census records, leading to better coverage of children aged 5 to 14
- better matching methods, which allowed the inclusion of additional records, linked between the NHS Patient Register (PR), the Department for Work and Pensions (DWP) Customer Information System (CIS), and Higher Education Statistics Agency (HESA)
- assigning records on the SPD to the most likely address using "activity" data from DWP benefit interactions and address moves recorded on the Personal Demographic Service (PDS)

Following feedback from users, we also improved the way we published the Research Outputs. We did this by using a new web-based format, improved data visualisation tools and a SlideShare presentation to help users understand the story behind the small area statistics.

Following the publication, users were asked to provide feedback, which will be summarised and published later in the year. Feedback from users included the following:

- SPD (v2.0) has shown an improvement in quality, which has the potential to be furthered by including "activity" data
- the production of population estimates for outputs areas (OA) would be useful
- suggestions that we could work with local authorities who have a detailed knowledge of their areas to improve estimates and understanding of the data
- generally, users were happy with the new release format

Users also suggested additional data sources that could improve the quality of particular population groups. These included the housing benefit register, electoral roll registrations and Driver and Vehicle Licensing Agency (DLVA) data.

What we're doing to improve the assessment

The next set of Administrative Data Census Research Outputs on the size of the population is due to be published later in 2017. We aim to respond to users' feedback, show developments to the methods, and produce outputs at output area level.

We will continue to explore the use of new 'activity' data for identifying and removing records from the SPD relating to individuals that are no longer usually resident in the population. This version of the SPD (v3.0) will be used in combination with a simulated Population Coverage Survey (PCS) drawn from 2011 Census data to produce coverage adjusted population estimates for 2011 by LA. This will take forward earlier research findings on adjusting for population coverage issues covered in [Beyond 2011](#). We are currently working with DWP and HMRC to acquire further data for this purpose.

We are also developing plans to test a PCS, to commence later in 2017. In the next three years, we plan to implement a comprehensive testing strategy to explore appropriate sampling and data collection methods, response rates and estimation methodologies. Our aim is to have a PCS in place by 2020.

6.3.2 Households and families estimates

| Evaluation criteria | 2016 assessment | Progress made since 2016 | 2017 assessment | Expected progress by 2018 | Expected assessment by 2023 |
|-----------------------------------|-----------------|--------------------------|-----------------|---------------------------|-----------------------------|
| Households and families estimates | RED/AMBER | ↗ | RED/AMBER | ↗ | AMBER/GREEN |

Evidence

The first set of Administrative Data Census Research Outputs on the [number of occupied addresses \(households\)](#) was published in February 2017. These examined the feasibility of producing household estimates from administrative sources. The Research Outputs provide estimates for the number of occupied addresses (households) at regional and local authority level for 2011, and compares these to the 2011 Census. They also provide estimates at regional level for 2015, making comparisons with the 2015 Labour Force Survey (LFS) household estimates.

The term “household” used for these outputs is based on the concept of occupied address, which is used in administrative sources. In contrast, census and LFS estimates, use a “shared facilities” definition of households, from which communal addresses are excluded. It’s unlikely that administrative data sources will be able to provide information based on the “shared facilities” definition. This is consistent with other countries that have moved to a register-based approach to census-taking, and we need to understand the impact of this for our users.

Following the publication users were asked to provide feedback on the outputs and the methods used. This feedback will be summarised and published later this year, and included the following:

- generally, users were content that the definition of “occupied address” satisfied their requirements for statistics on households
- as well as time series data, users expressed a need for households to be cross-tabulated with other characteristics such as income, employment, age and ethnicity
- suggestions on additional data sources that could potentially improve the quality of estimates included: Credit Reference Data, Council Tax data and electoral registrations

There is very limited information on families or relationships in administrative data sources. This provides challenges for producing household composition and family analysis based on ONS’s existing definitions. Alongside administrative data sources, survey information would be needed to derive household relationships and to meet the required definitions.

What we're doing to improve the assessment

We aim to improve these estimates by:

- developing methods to improve coverage, using a similar estimation approach to that being tested for the population estimates based on the PCS
- identifying and removing communal establishments from the list of occupied addresses using AddressBase removing communal establishments will help improve the coverage by removing addresses that shouldn't be included in our household estimates, and will help us adjust for some of the definitional differences described earlier
- seeking users' views on whether a definition of households based on 'occupied addresses' would meet user needs, including a jointly-hosted user workshop in July to understand the impact of the change in definition

We also intend to increase the breadth and depth of the "household" estimates in the next year, by publishing Administrative Data Research Outputs on the number of households for small areas (sub-local authority) and on the size and composition of "households".

6.3.3 Population, housing and household characteristics

| Evaluation criteria | 2016 assessment | Progress made since 2016 | 2017 assessment | Expected progress by 2018 | Expected assessment by 2023 |
|--|-----------------|--------------------------|-----------------|---------------------------|-----------------------------|
| Population and household characteristics | RED/AMBER | → | RED/AMBER | → | AMBER |
| Housing characteristics | AMBER | → | AMBER | → | GREEN |

Evidence

This year, the first Administrative Data Census Research Outputs on the local authority level individual gross income distributions were published for England and Wales. The report focused on the coverage and statistical quality of these initial income outputs with a view to improving this in future publications.

Users were asked for their feedback on the outputs, a brief summary of which is included below:

- household income would be useful for providing insight into living standards, financial inequalities and deprivation
- income by local authority and Lower Layer Super Output Area (LSOA) would be particularly useful, alongside income cross-tabulated with other characteristics such as employment status, tenure and household type
- for some areas (notably London), the top income band (£60,000.01plus) didn't reflect the variety of income in this category, as a notable percentage of the population earns over this threshold
- there is a user requirement for looking at variations in income over time

As mentioned in the earlier "Access to data" section, we've carried out an initial exploration into the availability and quality of administrative sources and their potential for producing outputs on population and housing characteristics. Figure 4 demonstrated that availability of administrative and other data sources in relation to census related topics is variable. The next step is to understand how good these data sources are and how they could be used to produce statistical outputs that meet user needs.

Assessing the quality of administrative data

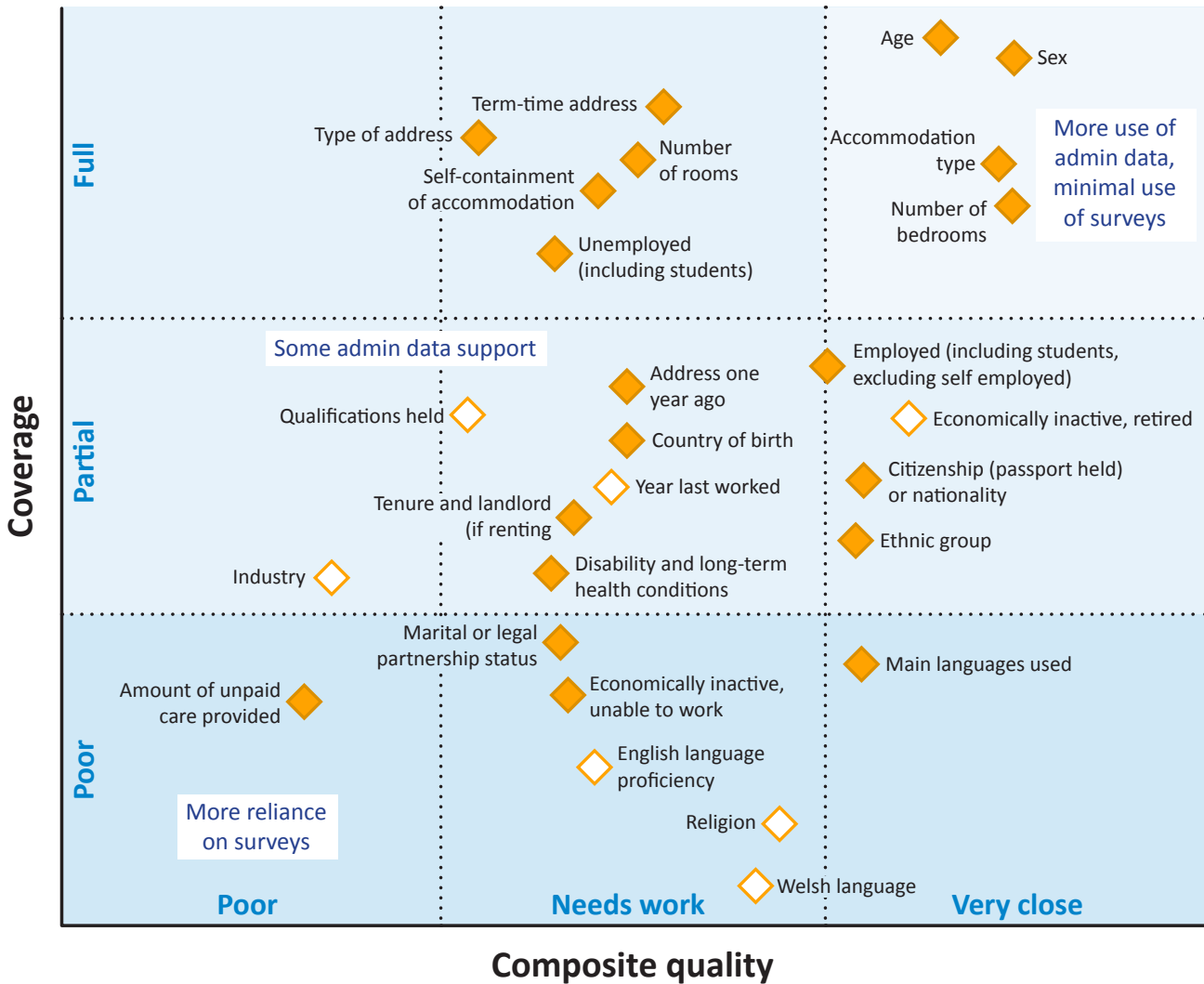
Our initial investigation into the potential for administrative data to produce information on population characteristics assessed each source against recognised quality dimensions, including: relevance, accuracy, coverage, timeliness and comparability. This was challenging due to the limited information available on many administrative data sources, particularly those we don't have access to.

The information we could find tended to be about the coverage, the data definitions, and possible errors in the data. We've used these key indicators to make a framework for measuring the potential for the administrative sources to produce characteristics outputs. This is shown in Figure 5, which provides a summary of our assessment.

This diagram simplifies a much more complicated picture. Each topic has its own issues and opportunities and there is a subjective element to scoring each source but it does provide a tool to show our progress now and in the future, as our knowledge grows.

The full quality assessment including further details on the scoring system and how the quality framework will be used in the future can be found in **Annex C**. We are interested to hear your views about how useful it is to present our findings in this way and would appreciate your [feedback](#).

Figure 5. Quality assessment of administrative sources



Key to chart

- ◆ Some data available to ONS
- ◇ Some data exists, but it is not available to ONS

Characteristics where data is not available, and there is insufficient meta-data to subjectively score:

- Household composition
- Number of hours worked
- Second residences (other addresses)
- Number of cars and vans
- Migration (Year of arrival, length of stay)

Characteristics with no accessible data:

- General health
- National identity
- Address of place of work
- Supervisor status
- Method of transport to place of work
- Economically inactive, looking after family
- Family relationships
- Occupation

The further a topic appears towards the top right of the diagram (full coverage, quality “very close”, lightest shading), the more we believe we can produce outputs based on administrative data alone, using survey data to help quality assure the estimates. The nearer the topic appears to the bottom left, the greater reliance we’ll have on survey data, with the consequence of less frequent and/or less detailed outputs.

The majority of topics will sit somewhere in between these two areas. These topics may be available from some administrative sources, but there is likely to be insufficient data to produce direct administrative-based estimates and there will be a need for further support from survey or other data sources. Where the topic measured in the administrative source is very close to the concept we’re trying to measure but there is only partial coverage (partial coverage, very close quality), then imputation methods will be required to fill gaps in the data. Another approach would be to use surveys to measure and adjust for coverage issues. Conversely, if the administrative data source has good coverage but imperfect quality (full coverage, quality “needs work”), an integrated approach with surveys could be used to adjust for this error. An example of this would be “general health”: information on health conditions which are collected across the health service could be used alongside survey responses of self-reported “general health” to provide a general health indicator.

Modelling approaches could be used to compensate for when coverage is imperfect, or for when the information available in the administrative sources is not of good composite quality (Full or partial coverage, quality poor or needs work). We have developed two types of model-based small area estimation methods: a regression approach and a structural approach. The first approach has been used to model unemployment estimates by combining unemployment benefit claimant counts from DWP with information on unemployment from the Labour Force Survey.

A model-based structural approach is useful if the administrative and other data sources have the same category structure and the administrative source has limited coverage. The generalised structure preserving estimation (GSPREE) approach is currently being tested to produce ethnic group population estimates.

There will be some characteristics traditionally provided by a census that can’t be obtained from administrative data. In this case, survey data alone would be needed to provide estimates.

The aim for an Administrative Data Census is to replicate as many census outputs as possible using administrative data and surveys. The idea is to produce these in a flexible way to produce aggregate cross tabulated estimates. This will be easier for those characteristics for which there are good quality administrative data sources available. For those characteristics with limited administrative sources, further estimation modelling and imputation methods will be required.

We will also be developing our requirements for a population characteristics survey (PCS) that will supplement the production of characteristics outputs produced by an Administrative Data Census. This work is being done alongside the transformation of ONS’s surveys as part of the [Data Collection Transformation Programme](#).

A further description of the methodological framework will be published later this year to accompany our future outputs.

What we're doing to improve the assessment

We are investigating the use of Valuation Office Agency (VOA) data to produce statistics on housing characteristics (number of rooms and bedrooms). The VOA data also contains information on age of property and floor space (topics not previously included in the census). The findings from this work will be published shortly. In the next year, we will be exploring the feasibility of producing Administrative Data Census Research Outputs from data that are currently available to us, on:

Census outputs

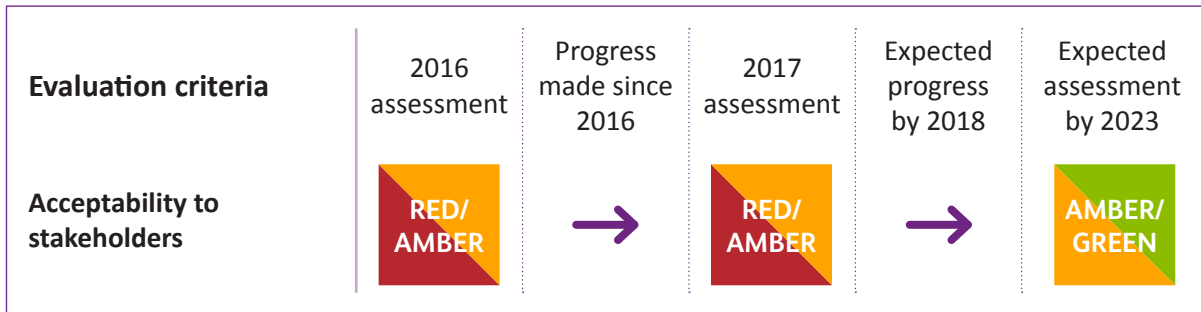
- ethnic group population estimates by local authority
- household size by local authority
- qualifications for school/university age population
- travel to work outputs using mobile phone data

New outputs

- further outputs about income (both household and personal income).
- mother's income in the (tax) year before birth

These new outputs are being produced to demonstrate the wider opportunities from using administrative data.

6.4 Acceptability to stakeholders



Evidence

Users

The assessment reflects the fact that until we can demonstrate that our methods can deliver outputs that meet the information needs of users, it will be challenging for an Administrative Data Census to be acceptable to stakeholders. We've demonstrated improvements in the quality of the population estimates, which have been well received by users, as has the production of a first set of research outputs on income and occupied addresses ("households").

We have engaged with a range of stakeholders about this work over the last year. This includes:

- a research conference held in June 2016 where we shared research updates, invited feedback on research plans and gauged confidence in our ability to move to an Administrative Data Census
- continued engagement with local authority and regional user groups with whom we shared research findings on our recent Administrative Data Census Research Outputs on population, "households", and "income" estimates. Feedback from these sessions is summarised in a report to be released later this year
- communicating progress through presentations delivered to, for example, the Government Statistical Service Methodology Symposium, British Society for Population Studies, RSS Stats User Forum, International Population Data Linkage Network, and European Conference on Quality in Official Statistics, amongst others
- biannual meetings that have been initiated between the devolved administrations to explore the use of administrative data in censuses
- representation on the United Nations Economic Commission for Europe (UNECE) taskforce on register based and combined censuses. Part of the taskforce's work is to agree on a set of guiding principles and a common framework for register based and combined censuses, which included ONS's ambition to move towards an Administrative Data Census
- working closely with countries who are also considering the potential for moving away from a traditional, 10-yearly census, and are doing similar research for example, Canada and New Zealand

Data suppliers

We are working closely with data suppliers to understand and improve statistical quality issues identified in the data. The Data Suppliers Group has been expanded to include members from key departments across Whitehall and the devolved administrations. This group is a forum for sharing research, discussing issues relating to data sharing, and building stronger relationships.

We've also set up statistical quality working groups with a number of supplier departments which meet several times a year. The aim of these working groups is to share knowledge on the quality of the data sources and to identify mutual benefits to improve the overall quality of the data.

Public

To safeguard the privacy of individuals, ONS have adopted the "five safes" framework to provide assurance that the data collected from individuals is only used for research using the following principles:

- data are handled by people who have been trained and accredited
- data are only used for research projects that deliver clear public benefits
- data are stored in a secure setting
- all outputs are checked and confirmed as non-disclosive before they are made available
- data are de-identified and have names, addresses and any other identifiable variables removed beforehand

The "five safes" are: safe people, safe project, safe settings, safe outputs, safe data. They are designed to address concerns raised by the public during research into the public acceptability of sharing their information. The common principles of the framework are used across government and academia in the UK and internationally.

The National Statistician's Data Ethics Advisory Committee (NSDEC) was established in 2015 to advise the National Statistician that the access, use and sharing of public data, for research and statistical purposes, is ethical and for the public good. The proposals for the personal "income" Administrative Data Census Research Outputs from combined Pay As You Earn (PAYE) and benefits data were considered by the committee in October 2016 before they were published in December.

Parliament

Parliament is likely to be concerned about the usability of the outputs, protecting the confidentiality and security of data, and public acceptability. Therefore all the evidence provided across the stakeholders is relevant. Additionally, the next section outlines what is being done to deliver value for money from an Administrative Data Census which is of high interest to Parliament.

What we're doing to improve the assessment

Users

We plan to further improve our methods and expand the range of outputs to demonstrate our ability to meet the information needs of users. We'll be submitting our methods and research for independent review later this year. The outcome from this review will be published.

Plans to further understand the requirements of users are underway. We are jointly hosting a user workshop in July on household information requirements. This will have a particular focus on understanding the impact of a change in household definition for users that would result from moving to an Administrative Data Census.

Data suppliers

The relationship forged through the Data Suppliers Group will be vital to unlocking data across government. Following the Digital Economy Act 2017, work will need to be done to put in place the practical arrangements for sharing data with approved safeguards.

Public

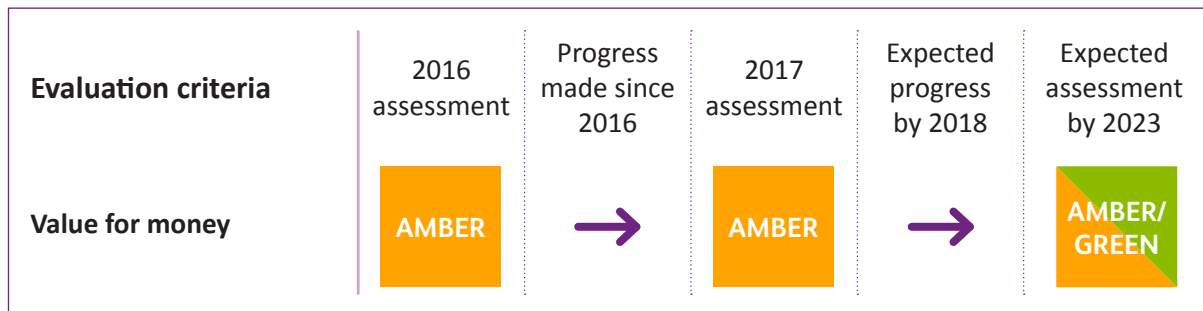
Future proposals for new Administrative Data Census Research Outputs will be put before NSDEC for guidance on the ethical considerations for the use of administrative data, in combination with survey and other data sources, where appropriate.

We will conduct further public acceptability research to understand the public's concerns around the use of their data. In 2023, a public consultation will gather views about an Administrative Data Census.

Parliament

As outlined previously, the work that is being done to improve the assessment with regards to users, data suppliers and the public should also improve the acceptability of an Administrative Data Census to Parliament.

6.5 Value for money



Evidence / What we're doing to improve the assessment

In the Beyond 2011 Programme, we published a [Summary of benefits of census information](#). We will be reviewing the benefits associated with moving to an Administrative Data Census.

7. Conclusion and next steps

This report assesses our current progress towards an Administrative Data Census. In it, we have described our plans to improve this assessment in the future, and we have identified the challenges we must overcome so that we can compare our outputs with those from the 2021 Census.

Despite these challenges, moving towards an Administrative Data Census offers a number of opportunities. These include producing key census outputs on a more timely and frequent basis, potentially for less money than the current system. It also offers opportunities to produce new outputs that aren't available through the current approach such as income, that could better meet user needs.

This year's assessment has focused on what we have been able to produce with the data we have. We've investigated the potential for administrative data sources to produce population, housing and household characteristics. Building on this work will gather pace as the new powers from the Digital Economy Act 2017 facilitate better access to the data sources we need. Next year's assessment will focus on the main uses of census-related information and how we are placed to produce outputs that meet these needs.

Over the next year, we'll:

- publish an expanded range of Administrative Data Research Outputs and seek feedback from users
- continue progress on acquiring administrative data and understanding the statistical quality of the data that are accessed
- submit our methods for External Assurance
- continue to engage with users and data suppliers through the Census Advisory Groups, Data Supplier Groups and other user working groups

Annex A. High level evaluation criteria explained

There are four key challenges to delivering an Administrative Data Census:

1. Accessing the range of data needed to produce outputs that are currently provided by the ten-yearly census
2. Linking together lots of independently collected data accurately whilst preserving the privacy and security of the data
3. Developing methods that can transform the linked data into outputs that meet the needs of users
4. Making an Administrative Data Census acceptable to key stakeholders, for example by providing value for money, and providing reassurance that data will be kept safe through this approach

To address these challenges, the following would need to be in place.

Rapid access to existing and new data sources

Criteria for assessment – access to data

To maximise the breadth and quality of statistics that could be provided by an Administrative Data Census, ONS would need to have rapid access to new and existing data sources from across government. This would also need to extend to other sources of existing data that would add value. ONS would also need to be consulted before changes are made to the administrative data that may affect the quality and stability of outputs from an Administrative Data Census over time. The Digital Economy Act 2017 offers a solution to these requirements.

The ability to link data efficiently and accurately

Criteria for assessment – ability to link

An Administrative Data Census would involve linking together multiple administrative data sources and surveys to produce statistics on the range of topics that the census currently includes. This isn't a simple task. Individuals in the UK don't have a single unique reference number that is carried across all government-held data, making this linkage challenging. For example, data about tax and benefits from DWP and HMRC use the National Insurance Number, while GP Register data uses NHS number and School Census uses a unique pupil reference number.

We need methods that enable us to link together these independent data sources accurately to enable the production of high quality statistics. An additional challenge is to do this while preserving the privacy and security of the data.

Methods to produce statistical outputs of sufficient quality that meet priority information needs of users

Criteria for assessment – ability to meet information needs of users

We need to deliver methods that can transform the linked administrative and survey data into statistical outputs that meet the priority information needs of users. This means providing statistics

on the topics users need, at the right level of detail (for example, for small areas), and at the right quality. In response to a public consultation in 2013, users told us we need to develop statistical methodologies that:

- provide robust estimates about the size of the population and the number of households
- provide estimates about population characteristics at a point in time to allow similar areas to be compared with one another
- provide the granularity of information that users need to measure change over time (for example being able to spot changes over a decade in unemployment rates by ethnicity for small areas)

Another important area, is developing the detail of the survey design that will be needed and the methods to model from surveys and administrative data.

Acceptability to stakeholders (users, suppliers, public and Parliament)

Criteria for assessment – acceptability to stakeholders

In order to successfully move to an Administrative Data Census in the next decade, users of the data, data suppliers, the public and Parliament need to be convinced that this approach meets their needs. Acceptability to the four main stakeholders (users, suppliers, public and Parliament) will be influenced by ensuring that:

- main information needs of users are met
- data are held, processed and linked while protecting privacy, confidentiality and security safeguards

Value for money – including benefits and costs

Criteria for assessment – Value for money

An Administrative Data Census will need to demonstrate that it provides value for money compared with a 10-yearly census. This means showing either that it can deliver the benefits that users get from a 10-yearly census at a lower cost, or that the cost saving is sufficient to justify lower benefits. For example the Administrative Data Census may not be able to deliver all the outputs that a 10-yearly census provides but it may include additional benefits such as more timely, frequent data and new outputs that are not currently provided by a 10-yearly census. This is the key trade-off that will need to be taken into account.

Annex B. Update on acquisition of data

This annex provides an update on data that we have obtained access to, and data that we are focusing on next.

[Data source overviews](#) have been published for a number of sources. These include statistical quality assessments which were carried out during the Beyond 2011 Programme on the key data sources which we now have access to. These reports covered primarily demographic and geographic variables with little characteristic information. The findings were written up as source reports and published on our website. The reports are:

- [Administrative Data Sources Report: NHS Patient Register \(S1\)](#)
- [Administrative Data Sources Report: Electoral Register \(S2\)](#)
- [Administrative Data Sources Report: The English School Census and the Welsh School Census \(S3\)](#)
- [Administrative Data Sources Report: Higher Education Statistics Agency: Student Record \(S4\)](#)
- [Administrative Data Sources Report: CIS combined \(S5\)](#)

Data sources that we are currently focusing on

We regularly review our priorities for the datasets we want to pursue and use. This means that some datasets referenced in the 2016 annual assessment publication may appear to have made little progress this year.

Personal level income and benefit information

Personal level income and benefit administrative data, if of sufficient statistical quality and legally accessible, will provide income, household and “activity” data. We have access to a subset of variables within these data from HM Revenue and Customs (HMRC) and the Department for Work and Pensions (DWP). This data includes some income and benefits variables for the population receiving benefits (including Universal Credit, Personal Independence Payments and Child Benefit) or tax credits, as well as information from the Pay As You Earn (PAYE) system (this excludes self-assessment and self-employment).

A source overview has been published for these datasets. The data has been used as an indicator of activity in Statistical Population Dataset (SPD) V2.0 and to produce the Income Research Outputs. We received an updated version of these datasets in December 2016: this will allow research to continue into how this data can be best used and enable us to refine our requirements for a more detailed data supply.

Access to the full data is a priority. The Digital Economy Act 2017 may help us gain access to the more detailed variables required and some specific subsets, such as self-assessment returns.

All education dataset for England

We are working with the Department for Education (DfE) to develop a longitudinal education dataset using the existing National Pupil Database, further education data, and higher education data (previously this project was owned by the Department of Business, Innovation and Skills). The dataset is expected to include variables on attainment and qualifications for everyone born from 1985 onwards through their secondary, further and higher education, as well as a range of socio-demographic variables. At present, coverage is for England only.

The data can potentially be used to improve estimates for specific population groups such as schoolchildren and students. This is possible using records as an indication of activity to determine where individuals are resident in the country, and their patterns of movement during period of study. We also expect the data to help us estimate the population by characteristics such as age, sex, ethnic group and qualifications held.

We compared a first version of this dataset (with a subset of variables) with the 2011 Census in order to assess the coverage and statistical quality of the matching used to create the data. The findings of this analysis have been shared with DfE and will inform further development of the dataset. A legal assessment is being completed to establish how the full dataset may be accessed. Our aim is to publish the first qualification research outputs in late 2017. This is dependent on accessing an initial data supply that is suitable for use.

Health data

We have access to demographic data for population statistics purposes, from health datasets held by NHS Digital. This data comes from the Personal Demographic Service (PDS).

A source overview was published on the PDS – movers' extract.. These data show all changes in postcode that took place within a set period of time. These data can be used as evidence of activity, from patients interacting with NHS systems through new registration or by updating their address or other details. This activity data was used within SPD V2.0.

We now also have access to a subset of the PDS "stock" extract, and contains similar information to the GP Patient Register. We're working with NHS Digital to understand what additional information can be provided through the PDS stock extract and to understand how it could be used in the future to improve our research outputs.

Access to "activity" and demographic data from Hospital Episode Statistics (such as ethnicity, but not including clinical information) is a priority for 2017.

Currently we have focused on health data for England only. We will need to expand this to include Welsh health data in the future.

Property attribute data

We have access to data from the Valuation Office Agency (VOA). This data contains variables on characteristics of address such as property type (for example detached, terrace) and size of property. We are currently reviewing this data to assess whether or not they can be used to produce estimates for the number of rooms and size of property.

Property website data

We have acquired two datasets from the Zoopla website about properties for sale or for rent:

- we bought cleaned Zoopla data for seven local authorities from WhenFresh, a data analytics company
- we acquired Zoopla data from the Urban Big Data Centre for the UK free of charge, but this dataset has only had minimal data cleaning undertaken

Using the property description from these datasets will add insight about properties, particularly those which are hard to count or access, such as gated communities or caravan homes (including whether the homes are more likely to be holiday or residential homes). As both datasets provide

details about properties over time, they may indicate the churn of people moving into or out of an area, which could help us improve population estimates.

Council tax information

The use of council tax data to indicate activity or the extent of population churn at an address is being considered. This data could provide evidence for an individual's location, which could be used to improve population research outputs. This year we have been working with a few local authorities to test the legal basis for sharing data. We've received council tax data from two local authorities and are currently assessing its potential use.

Mobile phone data

Mobile phones transmit data on their location back to mobile network providers. These location measurements have the potential to create a variety of estimates relating to population densities and flows. We've obtained a sample of commuting flows derived from the movement patterns of mobile phone users, and we're comparing them against Census travel to work estimates and other official data. The data provided to us only reveals commuting flows greater than 15 commuters so as to be non-disclosive.

Administrative data sources that we plan to pursue access to in the future

Feedback from the Census topic consultations and users has pointed to various administrative data sources for further investigation. These will be analysed and investigated based on their potential for producing outputs.

Further health data

Further health data collected by the NHS could provide additional activity information (such as receipt of prescription) and non-health characteristics (such as language) information. These could be used to improve the population and characteristics outputs.

Vehicle and driver data

We have held an initial requirements meeting with the Driver and Vehicle Licensing Agency (DVLA) to discuss accessing vehicle and driver administrative data. These data may allow the production of information about the number of cars or vans in households, which assists central and local government with transport and new housing planning. The data may also improve the administrative data population statistics research outputs, particularly among young males, who may be less likely to update their details on other administrative sources.

Electoral register

In 2013, Electoral register data was judged unsuitable for use as the sole source of information for the production of population and small area socio-demographic statistics. In June 2014 there was a move to individual electoral registration. We continue to work with the Electoral Commission to understand these changes and assess whether the electoral register data may be used in future.

TV Licence

Data about TV licences could provide additional information such as: churn at an address or "activity" data, when individuals transfer their licence, or information on addresses which may not be captured on other sources, such as caravan parks.

Other datasets for consideration

- Home Office data – exit checks
- Land Registry data
- Stamp Duty data from HMRC
- data to measure private sector housing tenure
- data to measure public sector housing tenure
- Royal Mail forwarding address data
- dwelling stocks and Department for Communities and Local Government (DCLG) and Welsh Government social housing returns
- electricity meter data
- business Valuation Office Agency (VOA) data
- destination of leavers from Higher Education Survey

Census topics by availability, and specific data sources

Looking at which data sources collect information on a particular topic is useful for targeting our efforts to access the right data sources. Some data sources could cover a range of population characteristics for example, ethnicity and religion are collected across education data.

The table below shows the data sources which are available for each census topic.

Please note: datasets in **bold** include more than one census topic

Table 1. Census topics by availability, and specific data sources

| Census topic | | Some data available to ONS | Some data available but ONS doesn't currently have access |
|--------------|-------------------------------------|---|---|
| Demographics | Household Composition | | Single Housing Benefit Extract (SHBE) |
| | Marital or Legal Partnership Status | Births, Marriages and Deaths Registers Single Housing Benefit Extract (SHBE): Lone Parent Indicator | Marriage Allowance (HMRC) |
| Education | Qualifications held | | All Education Dataset for England (AEDE) Higher Education Statistics Agency Student Record (HESA) Individualized Learner Record (ILR) The Universities & Colleges Admissions Service (UCAS) |
| | Term Time Address | Higher Education Statistics Agency Student Record (HESA) | All Education Dataset for England (AEDE) |

| Census topic | | Some data available to ONS | Some data available but ONS doesn't currently have access |
|--|--|---|---|
| Ethnicity, Identity, Language and Religion | Ethnic group | English and Welsh School Census Higher Education Statistics Agency Student Record (HESA) | Customer Information System (CIS) Hospital Episode Statistics (HES) Patient Demographic Service (PDS) Work and Pensions Longitudinal Study (WPLS) |
| | Citizenship (passport held) or Nationality | Migrant Worker Scan | Citizenship Data (Home Office, Immigration Statistics) Central Reference System (CRS) - Visa Data Higher Education Statistics Agency Student Record (HESA) Semaphore - Passport Applications |
| | Main languages used | English School Census Welsh School Census | |
| | English Language Proficiency | | Citizenship Data Individualized Learner Record (ILR) Patient Demographic Service (PDS) |
| | Welsh language | Customer Information System (CIS) - noted preference for Welsh documents | |
| | Religion | | Higher Education Statistics Agency Student Record (HESA) |
| Health | Amount of unpaid care provided | Carers Allowance - Department for Work and Pensions (DWP) | Hospital Episode Statistics (HES) |
| | Disability and Long Term Health Conditions | National Benefits Database (NBD) | Hospital Episode Statistics (HES) |
| Housing | Accommodation Type | Land Registry (LR) Valuation Office Agency (VOA) | Housing websites including Zoopla |
| | Number of Rooms | Valuation Office Agency (VOA) | Housing websites including Zoopla |
| | Number of Bedrooms | Valuation Office Agency (VOA) | Housing websites including Zoopla |
| | Second Residence | | Council tax data |
| | Self Containment of Accommodation | Valuation Office Agency (VOA) | |
| | Tenure and Landlord (if renting) | Single Housing Benefit Extract (SHBE) Valuation Office Agency (VOA) | Housing websites eg Zoopla |
| | Type of Address | Valuation Office Agency (VOA) | |

| Census topic | | Some data available to ONS | Some data available but ONS doesn't currently have access |
|---------------|--|--|--|
| Labour Market | Employed (including students, excluding self employed) | Pay As You Earn (PAYE) - HMRC | Higher Education Statistics Agency Student Record (HESA) Pay As You Earn (PAYE)- HMRC - to include pension earnings breakdown Self Assessment Data - HMRC (includes self employed) |
| | Hours Worked | | Pay As You Earn Data (PAYE) - HMRC |
| | Industry | Interdepartmental Business Register (IDBR) | Pay As You Earn (PAYE) - HMRC - IDBR linked to PAYE |
| | Economically Inactive, retired | National Benefits Database (NBD) | Pay As You Earn (PAYE) - HMRC - to include pension earnings breakdown |
| | Economically inactive, unable to work i.e. long term sick | National Benefits Database (NBD) | Universal Credit (DWP) |
| | Unemployed (including students) | National Benefits Database (NBD) | |
| | Year Last Worked | | Pay As You Earn (PAYE) - HMRC - to include pension earnings breakdown |
| Migration | Country of Birth | Birth registers | Central Reference System (CRS) - Visa Data Patient Demographic Service (PDS) |
| | Internal/ International Migration (including address one year ago) | | Central Reference System (CRS) - Visa Data Migrant Worker Scan Patient Demographic Service (PDS) Semaphore |
| Travel | Number of Cars/Vans | | Driving and Vehicle Licensing Agency (DVLA) |
| "Activity" | Interacting with a system from which data is taken | Child Benefit (HMRC) English and Welsh School Census Higher Education Statistics Agency Student Record (HESA) National Benefits Database (NBD) Patient Demographic Service (PDS) Pay As You Earn (PAYE) - HMRC Single Housing Benefit Extract (SHBE) Tax Credits (HMRC) | Department for Business, Energy and Industrial Strategy (BEIS)/Utility Companies Driving and Vehicle Licensing Agency (DVLA) Hospital Episode Statistics (HES) Individualised Learner Record (ILR) Personal Independence Payments (DWP) Self Assessment (DWP) Universal Credit (DWP) |

Annex C. A Quality Framework for Admin-based characteristics

The framework is built on the quality dimensions common to ONS outputs, which are based on those described by the United Nations Economic Commission for Europe (UNECE): Relevance, Accuracy, Timeliness, Accessibility, Interpretability, and Coherence. We have used these to draw out several “indicators”, which aim to capture the work and limitations involved in estimating characteristics from admin data. The dimensions and the corresponding indicators are described in Table 2.

Table 2. Current structure of the quality framework for Admin data-based estimates for characteristics

| Dimension | Indicator: Admin data application | Definition of Indicator |
|-------------------------------|---|---|
| Relevance | Data Definition | Does the definition provided by the admin data meet user needs? |
| Accuracy | Coverage | Does the available admin data capture the population of interest? |
| | Linkage | Can the admin data be linked? |
| | Errors in Data | To what extent are there errors within records: specifically, missing data and incorrect entries? |
| | Delays in updating data source | If individuals “join” an admin dataset late, or if data collectors are slow to “clean” individuals no longer in the population of interest from records, then the admin data may not accurately represent the target population for a characteristic. |
| Timeliness | Frequency | How often do we receive admin data from the data supplier? |
| | Time between event and available outputs | Time from when data are collected to when they are “ready-to use”: includes the time between collection and receipt by ONS, plus the time required to process the data and produce outputs. |
| Accessibility | Collected & available | Does admin data exist for this characteristic (is it collected)? Is existing admin data available to the ONS for this characteristic? |
| Interpretability ⁶ | Potentially relating to supporting metadata | This might include: Do we have the required information for interpreting the data correctly? |
| Coherence | Comparability over space | Does the quality of the other indicators vary over geography? |
| | Comparability over time | Does the quality of the other indicators vary over time? (e.g. data collection may be discontinued, replaced with new sources or changes in policy, according to the needs of services) |

⁶ This indicator is still not fully defined, which is a reflection of our being at a relatively early stage of exploration. We will define this indicator more fully in time.

The UK Statistics Authority has developed a [quality assurance toolkit](#) that requires users, producers and suppliers of administrative data to have a clear understanding of the quality of administrative sources. The toolkit creates a shared understanding across stakeholders from collection to dissemination and use. ONS is [developing guidance](#) on how to assess the quality of administrative data sources for a variety of specific purposes; we will develop our framework alongside this work, from collection to dissemination and use.

Developing the framework

Our approach is similar to that taken by Statistics New Zealand, who made [a framework](#) to describe their progress towards admin data-based estimates for characteristics in 2016. Key features include:

- The framework is for outputs about population and housing characteristics rather than individual data sources. So while individual data sources are considered, overall statements about quality are about the characteristics.
- Quality is scored for each indicator in a subjective way, relative to other characteristics (see 'How were characteristics scored?').
- Findings are visualised on a chart that compares coverage with a summary of the other quality indicators. This emphasis on coverage reflects the importance of this indicator when using admin data for estimation.

Evaluating the characteristics against the framework

Characteristics are scored for each indicator, on a scale of 1-5. The criteria used to score indicators for the current Assessment is provided at the end of this section.

Scoring criteria are based on two principles:

1. Objective definitions for 'poor' (1) and 'excellent' (5), where possible.
2. Comparison of indicators between characteristics, using subjective judgement (e.g. 'coverage is better overall for this characteristic than that characteristic').

The subjectivity of scoring reflects the limitations inherent to exploratory research: fully objective assessments of quality are not possible- or appropriate- at this early stage. Instead, the framework is pragmatic and useful for supporting and communicating our work as it develops.

We intend the framework to fulfil two roles: to describe our progress towards producing admin data-based estimates for characteristics, including a snapshot of where we are now; and as a tool to support future Admin Data Census research towards this goal- for example, by helping us to locate gaps in our knowledge and decide how to address them, resulting in better quality estimates.

Using the framework to assess ONS's progress towards an Administrative Data Census

Figure 6 provides an at-a-glance summary of quality across characteristics. The information available to us was not sufficient to support scores for every indicator in this figure. Instead, we have assessed the accessibility, accuracy (coverage, linkability and level of error), and relevance of the data sources for each characteristic.

Figure 6. Summary of quality across characteristics and dimensions

| Lower Quality | Quality | | | | | Higher Quality |
|--|------------------|-----------|-----------------|------------|---------------|----------------|
| | 1 | 2 | 3 | 4 | 5 | |
| Quality measures: dimensions and indicators | | | | | | |
| | Relevance | Accuracy | | Timeliness | Accessibility | |
| | Data definition* | Coverage* | Errors in data* | Linkage | | |
| Labour Market | | | | | | |
| Employed (non-students) | 5 | 3 | 4 | 5 | 3 | 5 |
| Employed (students) | 5 | 5 | 4 | 3 | 4 | 5 |
| Unemployed (non-students) | 4 | 4 | 5 | 5 | 5 | 5 |
| Unemployed (students) | 4 | 5 | 4 | 3 | 4 | 5 |
| Economically inactive: retired | 5 | 3 | 4 | 5 | 4 | 5 |
| Economically inactive: sick/disabled | 3 | 2 | 4 | 5 | 4 | 5 |
| Economically inactive: looking after family | 2 | 2 | 4 | 5 | 4 | 5 |
| Self-employed | 5 | 3 | 4 | 5 | 1 | 3 |
| Industry (NS-SEC) | 4 | n/a | 3 | 4 | 4 | 5 |
| Year last worked (NS-SEC) | 4 | 3 | 3 | n/a | 1 | 3 |
| Hours worked | 4 | 4 | 2 | n/a | 1 | 3 |
| Occupation (NS-SEC) | 2 | 1 | 1 | n/a | 3 | 4 |
| Supervisor status (NS-SEC) No admin data sources identified | | | | | | |
| Health | | | | | | |
| Disability and long-term health | 3 | 2 | 4 | 5 | 4 | 5 |
| Amount of unpaid care provided | 2 | 1 | 4 | 5 | 4 | 5 |
| General health | 2 | 2 | n/a | n/a | 1 | 2 |
| Religion | | | | | | |
| Religion | 5 | 1 | 3 | 3 | 4 | 3 |
| Education | | | | | | |
| Term-time address | 5 | 3 | 3 | 3 | 4 | 5 |
| Highest level of qualification | 4 | 2 | 3 | 3 | 4 | 3 |
| Housing | | | | | | |
| Accommodation type | 5 | 5 | 4 | 4 | 5 | 5 |
| Number of bedrooms | 5 | 5 | 4 | 4 | 5 | 5 |
| Number of rooms | 3 | 5 | 4 | 4 | 5 | 5 |
| Address type | 3 | 5 | 4 | 4 | 5 | 5 |
| Self-contained, or shared | 3 | 5 | 4 | 4 | 5 | 5 |
| Other addresses | 5 | 1 | 4 | 4 | 5 | 5 |
| Landlord | 1 | 5 | 4 | 4 | 4 | 5 |
| Tenure | 2 | 1 | 4 | 4 | 4 | 5 |
| Demographics | | | | | | |
| Marital status | 4 | 2 | 4 | n/a | 4 | 5 |
| Household composition | 2 | 1 | 2 | n/a | n/a | 3 |
| Household and family relationships No admin data sources identified | | | | | | |
| Ethnicity | | | | | | |
| Nationality/citizenship (passport held) | 5 | 3 | 4 | 4 | 3 | 5 |
| Ethnicity | 4 | 2 | 3 | 4 | 4 | 5 |
| National identity No admin data sources identified | | | | | | |
| Migration | | | | | | |
| Country of birth | 3 | 3 | 4 | 4 | 4 | 5 |
| Address one year ago | 3 | 3 | 4 | 4 | 4 | 5 |
| Migration: year of arrival | 3 | 3 | 4 | n/a | n/a | 3 |
| Intention to stay: short/long-term | 2 | 3 | 4 | n/a | n/a | 3 |
| Language | | | | | | |
| Main language | 5 | 2 | 4 | 5 | 3 | 5 |
| Knowledge of Welsh | 4 | 1 | 4 | 3 | 4 | 5 |
| English proficiency | 4 | 2 | 4 | 3 | 4 | 4 |
| Travel | | | | | | |
| Number of cars/vans owned | 3 | 5 | n/a | n/a | n/a | 3 |

Notes on table: Heatmap to indicate the degree of admin data support behind characteristics, according to scored quality indicators. Priority indicators are marked *. NS-SEC refers to “National Statistics Socio-Economic Classification”.

Developing the quality framework for future use

The framework is a useful way for us to demonstrate progress towards our twin goals of producing Administrative Data Census outputs for comparison with 2021 Census, and for producing new outputs.

It will provide a starting point for discussion with our stakeholders on the potential of admin data to produce characteristics outputs that better meet user needs. It will also help us to identify gaps in our knowledge or quality issues in the administrative sources, and to articulate these concerns with our data suppliers and users. Furthermore, the framework will allow us to identify potentially desirable trade-offs between the indicators - for example, permitting more error to enable more frequent outputs.

In the future we hope to explore the potential of combining admin data with other data sources to meet user needs. As our research advances we will require our quality framework to develop, so that it can continue to articulate our findings and so support discussion with users. Our intention is to develop the framework through its use- and by seeking feedback- into a pragmatic tool for supporting our progress towards a better Census.

Table 3. Scoring Chart for this year's Assessment

| Indicator (Dimension) | 1 | 2 | 3 | 4 | 5 |
|--|--|---|--|---|---|
| Data definition (Relevance) | Very poor or no match on dataset and therefore cannot answer the same question as the desired output | | Plausibly related variables e.g. subjective general health vs measured health | | Data answers exactly the same question as the respective census question or desired output |
| Coverage (Accuracy) | Population covered is not similar to the target population for the desired output at all | | Doesn't completely cover target population or over-covers target population eg exclusions like communal housing etc. | | The population covered in admin data is the same as the target population |
| Error in Data (Accuracy) | Lots of incorrect entries and missingness | | | | No 'mistakes' in data; no/few missing data |
| Linkage (Accuracy) | Cannot link to other data sources: no potential linking variables identified | | Variables that are likely to allow linking have been identified | | Have shown that linking can be done |
| Timeliness | Data has been received in a one-off dataset, with no regular updates | | Data is received regularly, but there is a large delay between collection and it becoming ready to use | | Data can be accessed by the ONS frequently, and with a short time between collection and being ready to use |
| Accessibility | No data sources identified, or data source identified but with no process underway to obtain it | | Data sources identified and in the process of obtaining | | The data is available in an appropriate digital format to be used by the ONS |

