The Census 2021 Data Asset longitudinal data source for population in England and Wales: design and plans

The Census 2021 Data Asset high-level design: exploring the feasibility of maintaining an anonymised person-level longitudinal data source based on Census 2021.

Contact:
Louisa Blackwell
pop.info@ons.gov.uk
+44 1329 444539

Release date:
27 January 2023

Next release:
To be announced

Table of contents

1. Main points

2. Overview of the Census 2021 Data Asset

3. Census 2021 Data Asset design

4. Proof of concept for Census 2021 Data Asset

5. Use of personal data

6. Potential to contribute to public good

7. Data security, confidentiality, legal and ethical considerations

8. Next steps

9. Cite this methodology

1 . Main points

- We are exploring the feasibility of maintaining an anonymised person-level longitudinal data source for England and Wales, based on data from Census 2021, which can be updated each year to reflect population change.
- We call this ambition for an anonymised person-level longitudinal data source the Census 2021 Data Asset (CDA).
- This article sets out the high-level design of the CDA and describes how we will start with a proof of concept stage to develop and test processes and methods, and also understand public acceptability of the CDA.
- The proof of concept will focus on how best to ensure the representativeness of the CDA as it updates through cohort maintenance: adding births and immigrants, and flagging deaths and emigration.
- The CDA will support longitudinal and cross-sectional analysis; it will also support cohort studies (satellite cohorts) based on population samples drawn from the CDA.
- The CDA and satellite cohort studies, taken together, have the potential to inform policy and decision making, and help provide greater and much more granular insight for planning and delivery of local services, housing needs and quality of life for the most vulnerable groups and for the population as a whole; this will provide unique insights into outcomes for different population groups based on their lived experiences.
- Examples of satellite cohorts include the Public Health Data Asset or studies on a specific population sub-group, such as the Refugee Integration Outcomes (RIO) cohort study.
- We treat the data that we hold with respect, keeping it secure and confidential, and we use methods that are professional, ethical and transparent only ever for statistical purposes and for the public good.

2 . Overview of the Census 2021 Data Asset

At the Office for National Statistics (ONS), we are exploring the feasibility of an anonymised person-level longitudinal data source for England and Wales, based on Census 2021 and then updated each year to reflect population change (births, deaths and migration); we call this the Census 2021 Data Asset (CDA).

The CDA forms part of our population and social statistics transformation programme, which aims to provide the best insights on population, migration and society using a range of data sources. In this article we set out our vision for the design of the CDA and how we plan to realise this through a proof of concept phase. This vision will form part of the evidence base for the 2023 National Statistician's Recommendation on the future of population and social statistics.

We aim to produce more timely, granular statistics to better meet user needs. At the heart of the future population statistics system will be a Dynamic Population Model (DPM), producing our best estimates of the population using a wide range of sources, including survey and administrative data.

To complement this aggregate-level modelled approach we intend to produce an anonymised person-level data source, which is representative of the usually resident population, with scope to extend to a wider set of definitions in future. This will support longitudinal and cross-sectional analysis, and cohort studies to address a wide range of policy needs. While the CDA will be a person-level data source, it will retain the ability to form households (based on Census 2021). Further research is needed to address how we maintain and update important demographic characteristics and also household-level relationships in the CDA.

We will set out our high-level view of the design for the transformed population and migration statistics system for England and Wales in a separate publication.

3 . Census 2021 Data Asset design

The Census 2021 Data Asset (CDA) aims to be an anonymised record-level representation of the usually resident population, with scope to extend to a wider set of definitions in future. Once developed, it will include demographics and characteristics such as age, sex, ethnic group, country of birth and nationality. It will be statistically controlled using weights from the Dynamic Population Model (DPM). That is, the population totals in the CDA will match those in the DPM by age, sex and local authority, ensuring coherence with published population totals.

Create a record-level base population for 21 March 2021

We start with record-level Census 2021 data before coverage estimation as the base population for 21 March 2021. This will exclude the 3% of the population that did not respond to the census.

We will account for this using records from administrative data such as the NHS Personal Demographic Service (PDS) data. The added records will match Census 2021 undercount characteristics and be confirmed as present in March 2021 by linking to electoral registers and education data, for example. These added records will have basic characteristics such as age, sex and local authority geography in the first instance, but we plan to explore how other main attributes such as ethnic group can be added through the planned proof of concept stage (Section 4: Proof of Concept for Census 2021 Data Asset).

The CDA is not a population register and will not be used for operational purposes, for example, to make contact with individuals (Section 5: Use of personal data). It will be an anonymised dataset with name information, date of birth and address removed.

The data and variables we plan to use for the CDA are listed. Data linkage and methods to attribute main characteristics will be developed during the proof of concept phase.

We will use personal identifying information only to link data sources. At this stage there will be minimal use of demographic attribute data from each source to report on linkage quality, population coverage and basic demographic analysis. We have identified the following datasets and preliminary variables and list their purpose in the CDA. We plan to use existing legal gateways to access the data we need to undertake the proof of concept.

Data sources and reasons for inclusion

We will be using a variety of different data sources to ensure the representativeness of the cohort is maintained over time. The following information outlines which data sources will be used and reasons for inclusion.

Census 2021 and Census Coverage Survey

- These will be used to form the population base.

NHS Digital Personal Demographic Service (PDS)

- This will be used to identify addresses and individuals missed by Census 2021.
- Additionally, monthly PDS update files will identify internal migration moves since Census 2021, new patient registrations from abroad, and emigrations.

England and Wales birth notifications and registrations

- This will be used to update the CDA with births since Census 2021.

England and Wales death registrations

- This will be used to update the CDA with flagged deaths since Census 2021.

England and Wales marriages and civil partnerships, and divorces and civil partnership dissolutions

- This will be used to address missed links because of name changes.

Home Office Exit Checks

- This will be used to update the CDA with new flagged immigration records since Census Day for EU and non-EU nationals.
- This will be subject to later confirmation after enough time has elapsed.

Home Office EU Settlement Scheme

- This will be used to flag the population with pre-settled or settled status.

Home Office citizenship

- This will be used to flag non-UK nationals who achieve citizenship.

Electoral registers for England and Wales

- This will be used to confirm residence as at March 2021 in England and Wales.
- Furthermore, overseas voters data will be used to flag British emigrants not identified in PDS data.

HM Revenue and Customs (HMRC) P85

- This will be used to flag emigrants, such as those who notified HMRC they are moving abroad.

English and Welsh School Censuses, Individual Lifelong Learning Records for England and Wales, and Higher Education Statistics Agency (HESA)

- These will be used to validate census records not linked to PDS where children or students are present.
- Additionally, these will be used to validate PDS records not linked to census where children or students are present.

Variables and reasons for inclusion

The following information outlines variables used to produce the CDA, the reason for inclusion, and which of the data sources they can be found in.

Full name

- Full name will be used in linkage and cohort maintenance.
- Can be found in all data sources outlined earlier.

Date of birth

- Date of birth will be used in linkage and cohort maintenance.
- Furthermore, date of birth will be used to derive age to report on linkage quality and analysis.
- Can be found in all data sources outlined earlier.

Sex

- Sex will be used in linkage, cohort maintenance, and to report on linkage quality and analysis.
- Can be found in all data sources outlined earlier.

Address (including postcode)

- Address will be used in linkage and cohort maintenance.
- Furthermore, this variable will be used to assign local authority to report on linkage quality and analysis.
- Can be found in all data sources outlined earlier.

Nationality

- Nationality will be used in linkage, cohort maintenance, and to report on linkage quality and analysis.
- Can be found in Exit Checks, census, and HESA data.

Country of birth

- Country of birth will be used to report on linkage quality and analysis.
- Can be found in death registration and census data.

Month and year of arrival

- Month and year of arrival will be used to filter data for linkage purposes, and to report on linkage quality.
- Can be found in census data.

Arrival and departure dates, and UK visa start and expiry dates

- Arrival and departure date, and UK visa start and expiry date will be used to filter the data for linkage purposes, and to report on linkage quality.
- Can be found in Exit Checks data.

Alternative addresses (including postcode)

- This includes usual address one year ago, second residence, and term-time addresses.
- Alternative addresses will be used for linkage and to report on linkage quality.
- Can be found in census data.

Term-time postcode or domicile address

- Term-time postcode or domicile address will be used in linkage and cohort maintenance.
- Can be found in HESA data.

Previous postcode

- Previous postcode will be used in linkage and cohort maintenance.
- Can be found in PDS data.

Ethnic group

- Ethnic group will be used to report on linkage quality and analysis.
- Can be found in census, School Census, HESA, and birth notification data.

NHS number

- NHS number will be used in linkage and cohort maintenance.
- Can be found in PDS and death registration data.

Date of NHS registration or date of patient UK entry

- Date of NHS registration or date of patient UK entry will be used to filter the data for linkage, and for cohort maintenance.
- Can be found in PDS data.

Reason for removal flag and other flags for new registrations

- The Reason for removal flag and other flags for new registrations will be used in cohort maintenance and to report on linkage quality.
- Can be found in PDS data.

Rolling the population forward

To roll the population forward to the mid-year, 30 June 2021, we will add records for births, confirmed long-term immigrants and then flag deaths and confirmed long-term emigrants to 30 June 2021.

For 2022, deaths occurring between 1 July 2021 and 30 June 2022 will be flagged in the CDA rolled forward population and new birth records added. As we cannot confirm immigration and emigration until 15 to 18 months after the event, we will use provisional immigration estimates from the DPM to identify and weight for new and returning immigrants based on age, sex and nationality.

For emigration we will apply weights based on estimates from the DPM by a core set of emigrant characteristics (for example, age, sex, nationality, geography). A cohort maintenance strategy will periodically confirm "candidate emigrants" using new admin-based activity indicating usual residence in England and Wales (Admin-data based migration estimates based on Home Office Exit Checks for non-EU nationals)) or confirmation of an emigration event through linkage to admin data sources such as Exit Checks for EU nationals, PDS, HMRC P85 or overseas voter data for British nationals, and HMRC Pay-As-You-Earn (PAYE) and Self-Assessment data for Irish nationals.

We aim to produce an annual version of the CDA and ultimately make this available to Office for National Statistics (ONS) accredited researchers in the ONS Secure Research Service environment (SRS) and the Integrated Data Service (IDS). This will be supported by regular cohort maintenance (the addition of weekly or monthly births and deaths, for example) ensuring that active records are not flagged as emigrants and proactively maintaining the integrity and representativeness of the CDA.

We will identify quality signals that demand corrective action, for example, a rise in linkage failure for deaths or births to mothers will alert us to under-coverage in the CDA. The planned proof of concept (Section 4: Proof of concept for Census 2021 Data Asset) for the CDA will test the feasibility of a maintenance strategy. The proof of concept will draw on lessons learnt from the maintenance of the Refugee Integration Outcomes (RIO) Cohort Study.

Census 2021 Data Asset design and maintenance

2021 base population at 21 March 2021 (Census Day)

- Start with census record level data necessary for linkage
- Include the main demographics based on Census 2021
- Add records missed by Census 2021 using administrative data sources
- Add new-borns missed by Census 2021
- Add Census 2021 visitors and short-term residents who may become future usual residents
- Flag emigrations, deaths and cross-border flows by linking administrative data sources

Cohort maintenance from 21 March onwards

- Add births
- Add new immigrations
- Add cross-border flows from Scotland and Northern Ireland
- Add other returning migrants
- Flag emigrations, deaths and cross-border flows by linking administrative data sources
- Then deal with residuals (unlinked deaths, emigrations, or births not linked to mothers), which will not retrospectively create a new record, but are flagged and added to linkage updates

Satellite cohorts examples drawn from the CDA base population

- Refugee sample: Home Office data
- Veterans sample: census and administrative data
- Migrants sample: census and administrative data
- In care and care leavers: census and administrative data
- Health Data Asset sample: Census 2021 Data Asset

Satellite cohorts

The CDA will support cohort studies based on population samples drawn from the CDA rolled forward population. Cohorts can be based on the entire England and Wales population, for example, the Public Health Data Asset (PHDA) or a specific population sub-group, for example, the Refugee Integration Outcome (RIO) cohort study. Examples of existing and potential cohorts follow.

The ONS Longitudinal Study for England and Wales

The ONS Longitudinal Study (LS) contains linked census and life events data for a 1% sample of the population of England and Wales. It contains records on more than 500,000 people usually resident in England and Wales at each point in time and is largely representative of the whole population.

The LS has now linked five successive censuses from 1971 onwards (Census 2021 data are being linked currently), allowing researchers to examine change between censuses. We are considering options to integrate the LS with the Census 2021 Data Asset (CDA) as a satellite cohort. The LS will uniquely enrich the CDA, for example, through:

- gold-standard record linkage, including clerical search and review for a core 1% sample that cannot be easily achieved by 100% linkage
- longitudinal history spanning six decades, against which new linkage for the CDA can be benchmarked and reviewed to ensure that the CDA is linking the same person over time
- providing early signal that the CDA record linkage algorithms may need adapting for hard-to-link population sub-groups

The CDA will offer the possibility of linking in additional attributes, for example, income and educational attainment for the whole population rather than a 1% sample.

Public Health Data Asset

The Public Health Data Asset (PHDA) linked 2011 Census with deaths and health data to provide timely evidence on the coronavirus (COVID-19) pandemic during 2020.

This Data Asset demonstrated the high value of linking events prospectively to census data to report on COVID-19 related deaths by ethnic group, disability and religion, but was limited as the base population was not replenished with births and immigration since the 2011 Census. We are planning future iterations of the PHDA where the underlying base population is the Census 2021 Data Asset that is replenished and rolled forward on an annual basis. The Census 2021 Data Asset will provide a means to add births and immigrants, and flag deaths and emigrants.

Refugee Integration Outcomes Cohort Study

The Refugee Integration Outcomes (RIO) Cohort Study is a collaboration between the Home Office and the ONS. The Study will provide unique insights into the integration outcomes for approximately 121,000 resettled and asylum refugees who were resettled under the Vulnerable Persons and Vulnerable Children's Resettlement Schemes or were granted asylum between 2015 and 2020.

The Study covers England and Wales currently, but there are plans to expand this Study to Scotland and Northern Ireland, and other humanitarian and protection routes in the future, subject to data quality, availability and funding.

Incorporation of studies such as RIO will strengthen the data inclusivity of the LS and CDA. We are using RIO as a case study to demonstrate how the development of this longitudinal cohort study can inform the development of the CDA. Specifically, we focus on the identification of migration events, cohort maintenance and loss to follow up, which are critical to maintaining the representativeness of the England and Wales population.

4 . Proof of concept for Census 2021 Data Asset

The Census 2021 Data Asset (CDA) includes a proof of concept development phase to demonstrate how the record-level population can be updated on an annual basis. This work is planned to start later in 2023. Ahead of this we have considered ethical standards and sought advice from the National Statistician's Data Ethics Advisory Committee (NSDEC).

The proof of concept draws on the Office for National Statistics (ONS) experience gained through the development of the Public Health Data Asset (PHDA). The PHDA was developed in response to the coronavirus (COVID-19) pandemic, providing insights to government and healthcare services to assist in the targeting of resources and messaging. The proof of concept will build on that experience, adding not just mortality but also other administrative data to a cohort that is defined by presence at Census 2021. Unlike the PHDA, the CDA will maintain the representativeness of this longitudinal asset through cohort maintenance: adding births and immigrants, and flagging deaths and emigration.

The proof of concept will develop and test processes and methods to create the CDA, including:

- reporting on quality and coverage of data to create the CDA and design of quality measures to understand the representativeness of the CDA
- exploration of data linkage methods and where clerical assessment of linked and unlinked records should be focused
- maintenance of a rolled forward population, with a focus on international migration using existing admin-data based migration estimates (ABME) data
- feasibility research into the maintenance and updating of important demographic characteristics and household level relationships using admin data
- exploration of how characteristics can be attached to records added through births and migration
- exploration of methods to produce uncertainty measures for a longitudinal data asset

5 . Use of personal data

The Census 2021 Data Asset (CDA) is not a population register and will not be used for operational purposes, for example, to contact individuals. Personal data (which for linkage include name, date of birth, address, or NHS number available on census and administrative data sources) are used for data linkage purposes only and will use the Reference Data Management Framework (RDMF). This is the Office for National Statistics's (ONS's) solution for the processing, management and linkage of data and will make sure that we maximise the benefit of the standardisation, transparency and consistency of data.

Personal identifying information will be removed from datasets created for analysis. All data processing takes place within a secure and tested processing environment that is used by the ONS for the majority of statistical processing of personal data. There is a separation of roles and project areas between those analysts who link the data and those analysts using anonymised linked and unlinked datasets for quality analysis. For example, comparisons of age and sex distributions in the linked records to England and Wales populations to check for representativeness.

We will ensure the risk of identifying individuals is minimised but balanced against utility of the data for analysis. Any published research based on the CDA will be at an aggregate level having been through strict disclosure control procedures. Demographic variables such as age, sex, ethnic group, country of birth or nationality will be added to the CDA for analysis purposes to support linkage but also analysis of the asset. These variables will be used for quality assurance analysis to assess the representativeness of the population across time and the inclusivity of marginal or vulnerable populations including migrants, refugees and those seeking or granted asylum.

Further information on data security, confidentiality, legal and ethical considerations for the CDA is given in Section 7: Data security, confidentiality, legal and ethical considerations.

6 . Potential to contribute to public good

The Census 2021 Data Asset (CDA) and satellite cohort studies, taken together, have the potential to inform policy and decision making, help provide greater and much more granular insight for planning and delivery of local services, housing needs and quality of life for the most vulnerable groups and for the population as a whole.

As an anonymised longitudinal person-level data source, the CDA can support both longitudinal and cross-sectional analysis, incorporating our best estimate of the usually resident population, with scope to extend to a wider set of definitions in future. Therefore improving on the Office for National Statistics (ONS) Public Health Data Asset [note 1] as a satellite cohort, through linkage of additional data to support analysis of life expectancy and healthy life expectancy estimates by characteristics and occupations not possible through the ONS Longitudinal Study (LS) [note 2], which is a 1% sample.

The CDA has potential to provide more granular analysis for the whole of England and Wales population than is currently possible through surveys and the LS. LS numbers quickly fall to single figures because of sample size when reporting longitudinal outcomes by age and ethnic group subnationally. For example, the ability to identify changes in morbidity and mortality at a local level across population groups, such as ethnic group, homeless people and children in care.

Satellite cohort studies, with samples drawn from the CDA and linkage of further attributes from administrative data sources can support analysis of educational attainment and health outcomes, currently not possible with the LS. Our aim is to ensure that the CDA builds capability to produce fully inclusive statistics on the whole population, not just those living in private households, but people living in, for example, communal establishments.

Notes for: Potential to contribute to public good

1. The Public Health Data Asset (PHDA) linked 2011 Census with deaths and health data to provide timely evidence on the pandemic during 2020.

2. The ONS Longitudinal Study (LS) contains linked census and life events data for a 1% sample of the population of England and Wales.

7 . Data security, confidentiality, legal and ethical considerations

We treat the data that we hold with respect, keeping it secure and confidential, and ensure that we comply with all relevant legislation, including the UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018.

For the Census 2021 Data Asset (CDA) we plan to use personal data such as name, date of birth and address to link data only. Other variables such as age, sex, ethnic group, country of birth and nationality will be used for analysis purposes.

We have sought assurance on the initial design of the CDA at the Office for National Statistics (ONS) assurance panels (Methodological Assurance Review Panel (MARP) and the Longitudinal Scientific Advisory Panel (LSAP)). We have also engaged with and sought advice from the National Statistician's Data Ethics Advisory Committee (NSDEC) on our proposal for a proof of concept CDA. We will continue to engage with NSDEC, LSAP and MARP throughout the development of the CDA and use this to inform a future decision on the viability of a full implementation of the CDA.

Personal data are only processed when necessary for statistical purposes. Only the minimum amount of personal data required to achieve the aim are used. A future CDA will only use a core set of data and variables for linkage, analysis and assessing the quality of linkages. Satellite cohorts will be based on samples drawn from the CDA where we only link the necessary data and variables needed to answer research questions. Both the CDA and satellite cohorts will use data and variables already held by the ONS. We will seek ethical approval for each satellite cohort as they are developed.

Personal data (which for linkage include name, date of birth, address, or NHS number available on census and administrative data sources) are used for data linkage purposes only and will use the Reference Data Management Framework (RDMF) and will be undertaken by the ONS within the secure data environment by experienced ONS analysts. There will be a clear separation between ONS analysts who access personal identifying data and those who access the resultant linked or unlinked anonymised data for quality analysis and reporting. Aggregate analysis will only be released from the secure data environment once disclosure control checks have confirmed that the risk of identifying individuals is minimised.

Anonymised data will be held within the ONS secure data environment for as long as they are needed for research and statistical purposes. Personal data used for data linkage will be held separately and kept for as long as necessary.

We will ensure ongoing legal and GDPR compliance during the development of the CDA. We have undertaken a Data Protection Impact Assessment (DPIA), which identifies and sets out mitigations for potential risks.

Our ambition is that the CDA and satellite cohorts become statistical research resources for use by ONS accredited researchers in the ONS Secure Research Service environment (SRS) and the Integrated Data Service (IDS). Accredited researchers will have access to an anonymised version of the CDA dataset or a satellite cohort for research and analytical purposes. Access to the CDA and satellite cohorts will be governed in a similar way to the ONS Longitudinal Study (LS) and SRS.

8 . Next steps

Public acceptability is essential to the development of the Census 2021 Data Asset (CDA), and we subscribe to best practice in public engagement.

The Office for National Statistics (ONS) is currently working with Health Data Research (HDA) UK, Administrative Data Research (ADR) UK and Public Engagement in Data Research Initiative (PEDRI) group partners to develop a collaborative public engagement campaign to improve public data perceptions. We will engage with the public, through focus groups and community engagement throughout the development of the CDA to ensure that it is developed and implemented for the public benefit. This approach will help the public's understanding of the use of data and give reassurance on how data are used and safeguarded. We will engage and be transparent with the public on our plans, listen to feedback and maintain an ongoing dialogue.

The ONS has an ongoing programme of work looking at public trust in the use of administrative data. We have been collecting views on the collection, storage and use of data for statistical purposes. We will continue to engage with community groups through the ONS Community Outreach team that supported Census 2021 and run focus groups to understand public acceptability of a CDA.

A report will be published on completion of the proof of concept for maintaining a person level representation of the England and Wales population. This report will outline the aims and outcomes of the proof of concept, details of the methodology (including data linkage quality criteria and the extent to which these were met) and high-level descriptive statistics on the characteristics of the linked and unlinked data. This work will be crucial to informing the development in the future, as well as the work of analysts undertaking data linkage as part of the ONS's Population and Migration Statistics Transformation programme. The report will be beneficial to analysts across the UK Government Statistical Service, the wider research community and inform international best practice on data linkage.

Provide feedback

We welcome your feedback on the Census 2021 Data Asset (CDA), our transformation journey, and our plans. If you would like to contact us, please email us at pop.info@ons.gov.uk.

You can also sign up to email alerts from the Office for National Statistics Population team for updates on our progress, and to hear about upcoming events and opportunities to share your views.

9 . Cite this methodology

Office for National Statistics (ONS), released 27 January 2023, ONS website, methodology article, The Census 2021 Data Asset longitudinal data source for population in England and Wales: design and plans