Methodology Advisory Research Panel

2021 Census Address Checking

1. 2020 Address Checking

This paper sets out the approach taken for clerical address resolution as part of the development of the 2021 Census Address Frame. The frame is used as the basis for making initial contact with both households and individuals in communal establishments.

ONS have undertaken undertaken a range of clerical activities as part of the broader work to understand and improve the quality of the frame. An accurate address frame is part of the ONS strategy of transforming how statistics are produced, with greater reliance on the use of administrative data. This paper is set on the context of this broader transformation.

2. Background

ONS use the AddressBase product as the basis for the census address frame. As described in the the annex, this is based on a range of different sources and is maintained by GeoPlace independently from the census operation. Epochs of AddressBase are released every 6-8 weeks. It is already used in a wide variety of applications across government and in the private sector. For use in the census, ONS supplement AddressBase with administrative data to ensure coverage of communal establishments.

An extract from AddressBase is used as the basis from the frame referencing June 2020. This is processed ahead of being delivered for initial contact letter in September 2020. A second extract from AddressBase, referencing October 2020 is then used to produce a ‘delta’ extract of new, case type changes which is loaded for printing in January 2021. ONS are in discussion with GeoPlace to identify new addresses which are added upto March 2021.

Address checking activity during 2020 was planned to (i) improve the quality of the frame and (ii) measure the quality of the frame overall. This involved a combination of physically visiting addresses and a clerical exercise using a range of desk based resources.

The process of checking an address is to make a judgement on whether a census contact letter should be provided. While a contact letter could be sent to evey building regardless of where it was a residential address, it would be very difficult to understand non-response patterns and so effectively follow-up non-responsding households. As well as being inefficienct such an approach would result in lower response overall.

As a direct impact of COVID19, ONS took the difficult decision to cancel the planned field address check planned for July/August 2020. In response, ONS implemented three key changes:

(1) Implemented a more strategic approach use of clerical resolution:

Adapted the Desk-based Address Resolution Team (DART) to analyse patterns highlighted by research of uncertain (or extract 3) addresses in collaboration with the Address Index team (both managed within Data Architecture division).

Introduced an operating model where analysis from samples of uncertain address types (or buckets) was reviewed with ONS subject matter experts and Census to decide whether address types should be included or excluded from the frame.

For example, DART work identified that for Houses of Multiple Occupation (HMO) there were 22,000 cases where there were individually addressed units such as bedsits. A decision was taken to treat these units as separate households (107,000 in total) given how unlikely it would be for a single household and for a form to be shared if treating as a single HMO. Where individual sub-units are not addressable, the HMO will be treated as a household (for example a shared student house).

(2) Supplemented clerical resolution with an automated data linkage approach

A bespoke data linkage process was been developed by Data Architecture to automate some of the resolution activity and to complement DART activity.

This linkage work used additional data sources to flag and resolve uncertain addresses reducing the burden of clerical resolution on the staff. The further development of the linkage solution for Delta is further shaped by the findings of the DART team which sits at the heart of the new address resolution operating model for Data Architecture.

(3) Increased resource working on clerical resolution

DART consisted of 17 core staff, with the addition of a further 19 supplied by Social Survey Division Staff. A separate Communal Establishment Address Resolution Team (CEART team) was stood up of 18 Social Survey Division Staff, this team was managed under Census.

3. Discussion

Our assessment of quality is that we are within the overall overcoverage target. This assessment is based on the initial frame delivery in August 2020, though DART resource has continued to undertake address resolution ahead of the delivery of the address delta.

A total of 195,855 records (through clerical and data linkage resolution). This compares to an estimate of 180,000 records when the decision to change the approach to address checking was taken. A total of 4,280 Communal Establishments were resolved (84% of the total considered) with a total of 405,267 addresses for individual rooms in student halls identified.

There are also some key learnings from the work which have changed our collection design. The main learning is about how we look for the patterns and

categories in each Epoch provided by AddressBase Premium (ABP). This knowledge and analysis of previous Epochs will allow us to continually improve the frame and monitor the pattern for each of the address categories.

Census Statistical Design, Data Architecture (Address Index and DART) and Census Field Operations collaborated closely throughout the work. Meeting weekly, the teams were able form a common view of priorities for resolution and to understand the implications for the collection. Working in this way has developed a deeper understanding of some of the most challenging address types.

It would be far more difficult to adapt and learn from findings during a field address check which took place over a far shorter six-week period. The cost saving has also been substantial.

Census and the wider transformation programme will benefit directly from the capability developed and improvements in quality. Specifically:

- Using the new address resolution operating model as the basis for an address resolution service to be used during the collection phase of the 2021 Census
- An operating model within Data Architecture which brings together data linkage, clerical resolution and subject matter experts to improve the quality of Address Base Premium (ABP) for the range of ways address data is used currently and as part of transformation plans.
- Data linkage methods to automate the checking of uncertain addresses when new epochs so clerical resource can be focussed on more challenging addresses.

4. List of Annexes

- Annex A: Address Resolution Approach and Findings

Annex A – Address Resolution Approach and Findings

Overview

This annex provides further detail on:

- Overview of the quality requirements for the address frame used in the 2021 Census,
- Process used in the creation of the frame,
- Design of the address resolution approach for under and overcoverage,
- Desk-based Address Resolution Team (DART) process and governance,
- Address resolution findings for households and communal establishments

Address Frame Quality Requirements

An accurate, high quality list of addresses is central to the statistical design of Census 2021 to ensure high quality statistics as well as protecting the ONS reputation.

As a result of the inevitable and continuous changes in residential addressing and with lack of a field address check to help sample the undercoverage and overcoverage of the frame, the team has developed processes to tackle these problems without leaving the Office. The step of assuring the quality of the frame is pivotal as it provides the framework to inviting households to participate in Census 2021.

Quality criteria were set out ahead of the compilation of the Census address frame and were dictated by the benchmark of Census 2011. Two factors were considered, namely

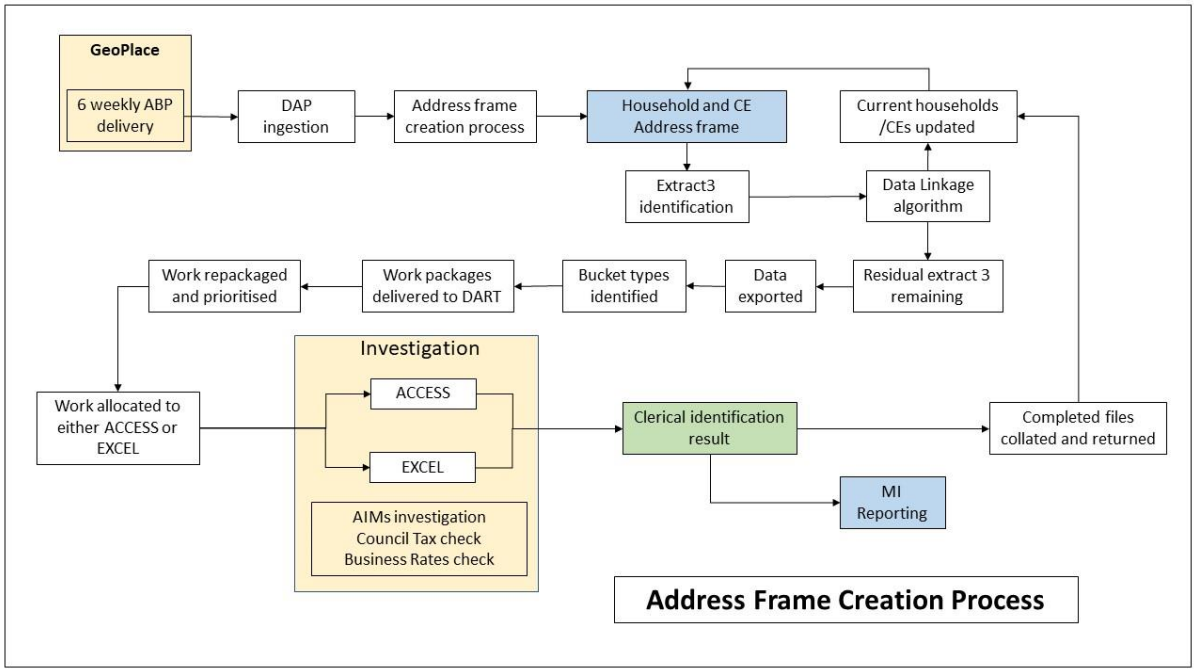
Undercoverage:

- Identifies 99.25% of residential addresses (no more than 0.75% under coverage).
- No more than 2% under coverage in any local authority.
- Identifies 100% of CEs with over 50 bed spaces.

Overcoverage:

- Less than 1% overall.
- Wrongly classify no more than 0.3% of addresses.
- Includes no more than 0.3% duplicates

Address Frame creation process



The frame is based on AddressBase Premium (ABP), a data source which is widely used across both the public and private sector and is maintained and updated by GeoPlace. Amongst others, ABP uses Local Land and Property Gazetteers (LLPGs) in conjunction with a range of address intelligence sources such as from the Valuation Office Agency (VOA), Royal Mail and Ordnance Survey to ensure the quality of the addresses it maintains.

ABP is provided on a six-week basis and the ingest and processing of this product has been managed in Data Architecture. GeoPlace have developed processes for understanding and assuring the quality of the address information and ONS hold regular calls with this data provider to understand the new releases, current quality issues and the next stage of development ABP provides the core of the Address Frame - however it is not perfect.

There are known limitations to the ABP product such as the quality of the addresses for the Communal Establishment or those associated with businesses. Additional data sources have been used to supplement the frame to identify missed addresses or misclassified shells as well as to validate the addresses of low confidence.

Supplementary sources include:

- Cushman and Wakefield (Student Halls)
- Care Quality Commission and Care Inspectorate Wales (Care Homes)
- Ministry of Defence and US Armed Forces (Armed Forces Bases)
- Ministry of Justice (Prisons)
- Edubase (Boarding Schools)
- Ministry of Communities and Local Government Survey of Traveller Sites

Address Resolution Design - Undercoverage

Undercoverage is usually estimated by a combination of administrative data analysis and address listing in the field. With the physical restrictions in place due to Covid 19 and the analysis of clustering performed on the PDS data, the decision was made to focus on reducing overcoverage through maximising the quality of address records that already appeared within the planned household frame.

In developing the design for the field address check a decision was taken to focus on overcoverage on the frame rather than undercoverage. This resulted in focusing on minimising the number of addresses which were either non-residential addresses or which no longer existed rather than addresses which were missed from the frame. The decision was based on:

- The nature of ABP as a resource that is maintained using established process and a range of sources used to provide timely updating of new addresses. No such mechanism has been used in any previous census.
- Research undertaken using GP Registration data (Patient Demographic Service – PDS) to search for clusters of missed addresses which could be resolved by clerical resolution. The PDS was independent from the sources used in ABP and has broad population coverage. No significant clustering was found.
- Findings from the 2019 Census rehearsal and recent address checking activity undertaken in Statistical Design and Research (SDR) division.

A field address check was carried out ahead of the 2011 Census which was targeted at the identification of new addresses. This used utility data to identify where there could be clusters of new addresses within postcodes potentially missed by the frame. Where a postcode was identified, field staff listed all residential properties. ONS are working with GeoPlace on whether new addresses identified on AddressBase from the delta extract to March 2021 can be identified and added to the census process.

Where addresses are not coverered by the frame and have not been added as part of a later update, households will be able to contact ONS to start a questionnaire. As part of the live address resolution ONS will look to identify patterns of new addresses (e.g. flats A, E and F).

Missed household addresses will also be estimated as part of the coverage estimation process. To be independent from the census, the Census Coverage Survey (CCS) involves an address listing exercises carried out by field staff rather than a predefined address frame.

Address Resolution Design – Overcoverage

ABP is already used in a number of areas of the ONS, as it used across government and the commercial sector. Within ONS ABP is used within the Address Index Matching Service (AIMS) to reference addresses on administrative data and as the basis for a range of business and social surveys (including the COVID Infection Survey).

On-going research to improve the accuracy of ABP for use in ONS identified approximately 500,000 'uncertain addresses' where for example an address was listed as residential but where no Council Tax record existed. These cases, referred to a 'extract 3' were classified into five categories:

- Multiple in Hierarchy
    - Addresses can be linked in what is known as a hierarchy e.g. a block of flats will have a primary address (address of main building) and then the flats themselves are considered children of that primary and be addressed by their numbers/letters.
    - Hierarchies where the primary is NOT extract 3 and there are 6 or more child addresses where one or more of these is extract 3
- Simple Postcode Clusters (5+)
    - Simple addresses are those that aren't linked to any others i.e. they are the only address in their hierarchy.
    - This bucket represents where there are 5 or more simple extract 3 addresses clustered by postcode
    - This bucket can potentially identify new housing developments
- Annexes
    - Addresses that have the word "ANNEXE" or a variation of that word in Address Line 1
- Misclassified Shells
    - The primary record in a hierarchy can also be known as a 'shell' record. These should be classified appropriately in ABP as such. Sometimes they can be mistakenly given residential classifications
    - This category consists of extract 3 addresses that have 3 or more child addresses associated
- Simple with PAF
    - Simple addresses that also have a PAF (Postal Address File – database maintained by Royal Mail) address linked
    - The implication here is that the property has received post

The planned field address check was designed to resolve 125,000 of these uncertain addresses, supplemented by a clerical resolution activity using the Desk-based Address Resolution Team (DART). This intelligence would also be used to provide an estimate of the overall frame quality. Clerical resolution included on-line searches of the latest published information on Council Tax and Non-Domestic Rates websites as well as internet searches for planning applications, sales and use.

National lockdown conditions during Spring 2020 placed restrictions on how ONS was able to undertake doorstep social surveys and a decision was taken to cancel the planned field address check. In response, ONS increased the size and scope of the DART work:

- Introduction of on-line recruitment and training. With all staff working from home, training was developed to ensure consistency of approach.
- Supplementing with additional Social Survey Division (SSD) staff. The 17 funded DART posts were supplemented by an additional 20 SSD staff who were unable to undertake doorstep interviews.
- Development of a data linkage approach. Using initial findings from clerical resolution, MDR and SDR held a joint hack day to identify ways to increase the number of addresses checked alongside the additional resource. As a result, an automated look up was developed to assess batches of the uncertain addresses using webscraped extracts of Council Tax and from the Valuation Office Agency (VOA).

Available SSD resource also provided the opportunity establish a Communal Establishment Address Resolution Team (CEART). Using 18 International Passenger Survey staff, CEART during April to August involved:

- Direct contact with University Halls to collect room level address information where this was not available on AddressBase or Cushman & Wakefield data.

- Direct contact and online searches to check the classification and capacity of a range of other CE types on AddressBase where administrative data was not available (including religious establishments, boarding schools, and nursing accommodation).

DART Process and Governance

Using the knowledge that already existed within the addressing team, training was developed for each bucket type to target particular outcomes.

The extraction of the buckets was carried out by the AI team and records were passed for resolution to DART. Resolved records were managed within the DART and returned to the AI team who used the DART outcomes to inform the frame design.

The allocation of bucket types was particularly important where the DART utilised the SSD workforce. Less complex cases were provided to these less experienced staff (such as under construction), enabling the core DART team to focus on the more complex cases while still providing a quality assurance role of the less complex cases.

Training was organised through the use of 'all day conference' calls. This innovative program ensured consistency of training within a shared 'safe' space. All trainees benefitted from questions raised by individuals, in a similar way to a standard classroom set-up. This proved to be an extremely supportive team environment which continues throughout all DART work with all day conference calls continuing.

The work progress had been communicated through governance meetings on a weekly basis up to the delivery of the initial frame at the end of August. The strategy for supplying the Delta was agreed and DART are now committed to providing updates regarding progress on a 3-weekly basis with a more detailed update on how the strategy may evolve as more data analysis is carried out relating to DART outcomes, data linkage and epoch on epoch changes.

Address Resolution Findings

Household

The volume of work undertaken by DART and the associated linkage work is summarised as:

- Total number of clerical records checked by end August = 207,108
- Total number of clerical records checked since end August = 55,987
- Total number of records resolved through data linkage = 73,319

Not all cases clerically resolved by end August will be used in the Address Frame. It was necessary to use an earlier version of Address Base Premium (e73) to identify uncertain cases as e77 was only available in July. Between e73 and e77 a number of cases were corrected in the sources data so did not need to be resolved – it is not possible to identify which cases will subsequently self-resolve. The total number of clerically resolved cases used in the Address Frame was 136,652.

Since the end-August delivery, DART have resolved a further 55,987 cases.

A summary of the findings from DART and data linkage are provided in table 1 by bucket type. As planned not all records were resolved. By end-August there were 176,810 unresolved cases from the buckets considered, alongside 139,393 cases which were uncertain that were not in a bucket (as they had other reasons for uncertainty).

We can use the cases which were resolved to estimate how many of the remaining outstanding cases are false addresses so will be sent a unique access code in error. This is estimated to 73,218 cases. If we assume that all uncertain cases outside buckets are false addresses (139,393) then the estimated total number of false addresses is 212,611 or **approximately an overcount level of 0.81%.** As not all cases outside buckets will be false, this will be an overestimate and further work is being undertaken to evaluate.

Table 1 – DART and Data Resolution to end-August

Bucket	Number of Records Checked by DART	DART proportion of current residential addresses from sampled bucket	Resolved by data linkage	Number outstanding in the bucket not evaluated by DART or resolved by linkage	Estimated number of current addresses for those not evaluated by DART	Estimated number of false addresses for those not evaluated by DART
Multiple in Hierarchy	72,078	40.50%	10,625	105,842	42,866	62,976
Simple Postcode Clusters (5+)	48,838	90.30%	2,660	18,060	16,308	1,752
Annexes	6,322	8.60%	51	745	64	681
Misclassified Shells	1,123	16.10%	24	54	9	45
Simple with PAF	4,875	85.10%	39,483	52,109	44,345	7,764
Not in a Bucket	3,416	N/A	20,476	139,393	N/A	N/A
Total	**136,652**		**73,319**	**316,203**	**103,592**	**73,218**

A summary of the DART work undertaken since end-August is set out in Table 2. This reduces the number of estimated false addresses within buckets to 40,422.

Table 2 – DART cases resolved since end-August

Bucket	Number of Records Checked by DART	DART proportion of current residential addresses from sampled bucket	Resolved by data linkage	Number outstanding in the bucket not evaluated by DART or resolved by linkage	Estimated number of current addresses for those not evaluated by DART	Estimated number of false addresses for those not evaluated by DART
Hierarchy reviewed since delivery	55,987	62.40%				
Total hierarchy reviewed	128,065	50.10%		60,480	30300	30,180
Total	**192,639**			**260,216**	**91,026**	**40,422**

Communal Establishment

The bulk of Communal Establishments were identified using the AddressBase Premium classification code system supplemented by a variety of admin data sources. Research was carried out by the Address Index team to determine useful admin data sets and the data was acquired. The data was then matched to ABP using the AIMS tool and UPRN assigned, enabling the special census classification variables ('ESTAB_TYPE and 'ADDRESS_TYPE') be assigned to true address records with confidence even where uncertainty existed around the ABP classification. A series of CE frames were created for each' ESTAB_TYPE' and these formed the basis for the CE list in the final frame. From this process

we were able to identify further requirements such as missing University Halls units, and feed this into the CEART clerical resolution process.

All Communal Establishment clerical resolution activities were also completed by end-August.
The work consisted of
(i) acquiring room level address (student halls of residence) and
(ii) checking the current status of other CEs and acquiring bedspace information
(approved premises, boarding schools, other educational establishments, residential children homes, hostels, low/medium secure mental health institutions, religious communities and staff accommodation).

As summarised in table 3, room level address information was collected for 81% of halls where this information was missing. A total of 405,320 room level addresses were acquired. Where further investigation of other CE types was required, 87% of cases were resolved.

It was not possible to identify information for all CEs. Not all CEs could be contacted or were able to provide the required detail. In some cases it was not possible to confirm current status or bed space information through on-line searches.

Table 3 CEART Resolved Cases

	Required Resolution	Resolved	Percentage
Student Halls	2288	1846	81%
Other CE	2808	2438	87%
All CE Types	5096	4284	84%

1. Annex - DART QA Overview – 01/12/2021

Objective – In response to MARP comments this overview sets out the Desk Based Address Resolution (DART) quality assurance process for where uncertain addresses were resolves through clerical resolution.

1.1 Workplace Address QA E2E

1.1.1 Stage 1 – Pre - Resolution

- On receipt of files a unique DART ID Reference Number is allocated to each record. The ID will stay with the record for the duration of its time in DART process
- This original complete list with DART_ID added is then stored separately from the processing work. This is then used at the end of resolution and collection as reference to ensure all records have been through the process and returned
- The files are then divided into batches of spreadsheets, which, in turn, are added to a tracker. The tracker is updated by our Data Delivery Manager and shows the status of all work at any given time

1.1.2 Stage 2 – Processing Run 1

- Once records have been clerically reviewed once we initially check that there are no blank records that are returned with no outcome.
- We then check for records that have more than one outcome as this indicates uncertainty. For this work, it can only have an address or unresolved reason, not both. Where there is an address string and an unresolved reason, the unresolved reason is removed. Where there is a missing outcome, it can be worked by the person completing the QA or if there are more than a couple – sent back to the person who completed the work so that they can review/amend
- Finally, we check for any data type conversions to make sure no part of the data has become corrupt. All records that have passed the above checks are then appended to a central master file.

1.1.3 Stage 3 – Outcome sense check and Run 2

- Where the record has been resolved in run 1 (in this case, there is an address string with a postcode), it is run through the automated Address Index Matching Service (AIMS) to check the validity of the address/postcode
- For records marked as unresolved – we carry out a sense check to ensure the correct unresolved status has been applied
- Where we identify records that could be resolved with more information/time, if not already marked as ‘Address Not Found’ the outcome of the record is altered to ‘Address Not Found’
- At this stage the master file is split – records that have passed all checks remain in the master – all files now marked as ‘Address Not Found’ are re-issued to DART for a run 2
- Following run 2 – the records are put through Stage 2 and 3 QA again (maximum twice run through) and then added back into the master

1.1.4 Stage 4 – Final check of master file

- The master file is added to Access and DART_ID is set as a key field to prevent duplication
- Queries within Access then check to ensure the records that were received are contained in the master file so that none are missed.
- Totals are checked against the original file to ensure everything tallies and the original data is matched with the original reference numbers that we received.

1.1.5

1.1.6 Areas improved over time

- DART ID Reference number – The reference numbers are now allocated on receipt of the file before anything else is done, with a copy saved separately for comparison throughout – previously this has sometimes been applied once files have already been batched or cleaned. By doing this, we have added confidence in several of the checks throughout the stages above
- Tracker – A tracker has been added so that we can confidently monitor every batch/spreadsheet that we allocate, ensuring no work is lost in the process
- Targeted Run 2 – Due to specific data types/volumes and delivery times, we now target run 2s for maximum resolution that fits requirements rather than generic 2nd runs to ensure quality.
- Improvements to stage 4 – We now add a check to ensure that the reference and data we receive matches the same data and reference going back (Ref 1/Shop A doesn't go back as Ref 1/Warehouse B) previously, we only checked that reference numbers weren't duplicated