NON ROYAL MAIL IP OPEN ADDRESS REGISTER: PILOT FINAL REPORT

RESPONSIBILITY FOR THIS DOCUMENT

[redacted] are responsible for the content of this document.

PUBLICATION DATE

Version 1.0 - September 2016

CONTENTS

1. EXECUTIVE SUMMARY

Ordnance Survey (OS) was commissioned by the Open Address Register (OAR) Project Steering Board to explore options for the creation and maintenance of an address register which contains no intellectual property rights of Royal Mail Group Limited. OS prototyped the methodology outlined in the Non RM-IP Address Register Solution paper, previously submitted to Steering Board, on a significant sample of properties covering a range of address types in a variety of geographies.

OS identified **32.5 million candidate records** for an address, and used three methodologies to populate the address for these records: Fully automated based on existing OS content; Desk based manual analysis; and Field based data capture.

The results of the prototype, when extrapolated up to a national scale, are:

Completeness in comparison to candidate list: **96% (1.3 million incomplete records)**

Accuracy[1] of completed OAR product content: **90.8% (2.9 million incorrect addresses)**

Thus, 4.2 million (13%) addresses that are present within AddressBase would either not be in the OAR, or would be in erroneously.

Several trends have been identified with regards to the OAR completeness and accuracy:

High levels were achieved where:

- There was one address within a single property (a one-to-one address), and the name / number had already been captured by OS in a non-address product
- A one-to-one address, and the number could be interpolated from nearby addresses where the number had already been captured by OS
- An address was an Object Without Postal Address[2] (OWPA), e.g. a bill-board / electricity sub-station
- An address was unique to Local Authority and not captured by Royal Mail

Lower levels were achieved where:

- There was more than one address within a single property (multiple-to-one addresses), e.g. a block of flats / shopping centre
- The address was accessed from a private shared drive (e.g. a block of flats / industrial estate)
- The address only had a name
- The address was uniquely identified only by an organisation name

These trends are primarily down to the existing specifications of the key OS datasets used in the Pilot: AddressBase, OS MasterMap Topography Layer and OS MasterMap Integrated Transport Network (ITN) Layer. The Pilot has used these products for purposes that they were not designed for, particularly within multiple-to-one addresses properties. Increased investment in the capture of the aforementioned OS datasets could improve the accuracy and completeness of the Pilot OAR dataset, for example improving the internal capture and representation of the properties that contain multiple-to-one addresses, e.g. shopping centres / train stations. This would overcome many of the data issues encountered in this project. Further work would be needed to determine the required investment and quality impact on the OAR data as this was out of scope for this Pilot.

This Pilot has also demonstrated that there is a critical issue is the linking of multiple-to-one addresses with the existing Unique Property Reference Number (UPRN). The UPRN has been accepted and adopted across Local and Central Government as the key for linking address datasets.

[1] A description of how the accuracy was calculated can be found within section 2.5 Quality Analysis Method Overview

[2] Objects Without a Postal Address (OWPAs) – These are records which are captured by Local Authorities due to their extended business requirements when compared to Royal Mail. Therefore, include records which attract rates such as Advertising Hoardings. These records also include items such as Ponds and Electricity Sub Stations.

The only methodology available for appending a UPRN to an OAR address, results in a potential UPRN error in up to 18% (5.9m) of OAR records, thus making the continued use of the current UPRN unviable.

It should also be noted that whilst the results of the prototype when extrapolated up to a national scale were encouraging, at 96% completeness and 90.8% accuracy, the following should be considered:

1. 96% completeness does not provide complete coverage of GB – for example 1.3 million addresses would be incomplete and therefore not included within the register
2. 90.8% accuracy of the address register would result in approximately 2.9 million complete addresses being incorrect.
3. These 4.2m could not be identified prior to the product release, without significant quality assurance investment, such as checking large volumes of records manually. Without this, confidence in the product would be seriously undermined by over 1 in 10 addresses being missing / erroneous, affecting usability, especially by organisations such as emergency services and other customers where accuracy is key.

From the Pilot it has been demonstrated that the above results could be achieved in 5 ½ years, at a cost of [redacted] with ongoing maintenance fees of [redacted]

Based upon the evidence contained within this paper, it is the recommendation of OS that [redacted] this OAR Pilot is not taken forward as a potential solution / negotiating tool with Royal Mail. This recommendation is based upon:

1) The lack of a postcode, required for uptake by citizens and users (both public & private organisations)
2) The completeness levels (96%) resulting in an incomplete address dataset
3) The accuracy levels of the completed records (90.8%) resulting in too many errors (2.9 million) to be accepted as definitive / usable by customers for their business applications
4) The existing UPRN would no longer be viable
5) Significant confusion in the marketplace due to the creation of a second spatial address dataset – something that was overcome, after many years of consternation, by the creation of AddressBase.

2. PILOT PROJECT

2.1 BACKGROUND

In Budget 2016, the government made a commitment to explore options for the creation of an open and free to use address register. The Project Steering Board have asked OS to explore options for the creation and maintenance of an address register which does not contain any intellectual property rights of Royal Mail Group Limited. OS proposed in outline, how such a register could be created and maintained within the solution paper "Non RM IP Address Register". Subsequently OS was commissioned to undertake a Pilot project to ascertain further detail concerning a potential solution containing no Royal Mail IP, in particular around quality, costs and timescales. OS agreed to undertake such a Pilot project.

2.2 AIMS

The Pilot prototyped the methodology outlined in the Non RM-IP Address Register Solution paper. Evidence of the address quality (completeness and accuracy) for each stage of the production methodology was demonstrated.

The Pilot focussed on a substantial sample of properties covering a range of address types in a variety of geographies. This was to ensure that any issues associated with different address types and geographies were encountered.

The methodology that was tested is outlined under the section 2.4.

The cost estimates to create and maintain the solution were validated and updated from the findings of the Pilot.

2.3 ASSUMPTIONS

The parties agreed that the Pilot would be undertaken with the following assumptions:

- The quality of the data developed as part of the Pilot project should attempt to match the current AddressBase offering for completeness, accuracy, and currency for the fields within the scope of the Pilot project, **but must not contain any RM IP**.
- [redacted]
- [redacted]

Third party data would not be used in the initial capture, although such data could be investigated for its potential to be used in the creation and maintenance of an OAR.

[redacted]

The complete OAR must cover addressable properties[3] within Great Britain. **To be clear, the OAR would not contain a postcode.**

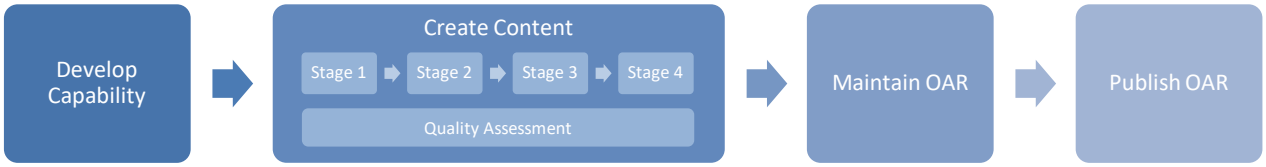
As a minimum, every record in the OAR must have the following primary fields:

- UPRN (every addressable property)
- USRN
- Building name / Building number / Occupier (if there is no other way of identifying)
- Sub building name / number (applicable for flats)
- Street Name
- Additional addressing content (locality, town, county)
- X, Y coordinates

[3] An addressable property in this project is defined as anything that attracts a rateable value. It would include electricity sub stations and billboards but not geographical features such as ponds or cattle grids.

2.4 CREATION & MAINTENANCE PROCESS OVERVIEW

The following phases of activity would be required to create, maintain and publish an address register. The Pilot Project primarily focussed on testing the methodology to create OAR content, but has also explored the required activity and cost associated with developing capability, maintenance and publication of an OAR. Each of these phases are described in more detail within the report.



The high level methodology for the "Create Content" phase is as follows:

STAGE 1 **Generate Candidate List**	➢ Creation of an address candidate list from the NAG which contained the UPRN, USRN and X/Y coordinate
STAGE 2 **Automated geo-processing**	➢ Sophisticated geo-processing to identify the building name/number from OS data ➢ Use of OS products to assign additional address information (street, town, ward, county)
STAGE 3 **Desk based geo-processing**	➢ Manual desk based intervention in scenarios where automated processing could not identify a building name / number or locality ➢ Use of OS data and investigation of other third party data sets for manual validation and data enhancement
STAGE 4 **Field geo-processing**	➢ Manual field based geo-processing throughout GB of remaining address candidates or part created addresses
Quality Assessment	➢ Continued quality assessments through each stage of creation

2.5 QUALITY ANALYSIS METHOD OVERVIEW

Reporting on the quality of addresses at each production stage was essential for the success of this Pilot. Both the completeness and accuracy of the OAR were measured, at various levels of detail. An outline of the approaches taken to measure quality is as follows:

- **Completeness**

  There are over 20 fields within the sample specification, but not all of these fields needed to be populated in order to create a "complete" address. The essential primary fields required to make a complete address were identified. Please see section 2.3 for the list of primary fields.

- **Accuracy**

  To determine record accuracy, two key steps were taken:

  1) **The OAR addresses were compared with AddressBase**
  The address text string was matched using the UPRN and any difference in characters recorded. For an address to be a valid "match" the building number had to be exactly the same between AddressBase and OAR; and the rest of the address had to be less than 6 characters different. For example:

  OAR: 23 Saint Oswald's Close, Kettering
  AddressBase: 23 St. Oswalds Close, Kettering

Analysis has shown that a threshold difference of less than 6 characters accounted for allowable differences between addresses (abbreviation or grammar).

In this way, the AddressBase product has been used as a reference dataset. It is important to note that AddressBase was only used a reference dataset as part of the quality assessment process. It is also important to note that in any implementation of an OAR creation it would not be possible to compare the addresses created against AddressBase due to RM IP implications.

2) **Manual analysis of records that did not "match" AddressBase**
In circumstances where the OAR was found to not match AddressBase (more than 5 characters different), desk based analysis was conducted to assess whether the address was still valid. Editors compared the OAR and AddressBase records and visually checked where the difference in the address string was. If required, OS data sources were used to support the analysis to establish whether an address was different from AddressBase and still valid, or incorrect. Examples of the analysis are reported under the accuracy findings for each content creation stage (Sections 4.2.2, 4.3.2, 4.4.2).

- **Address element analysis**

  Address quality was measured at several levels including the individual primary fields, by comparing these to AddressBase. This identified common trends and enables the OAR project team to act on the findings.

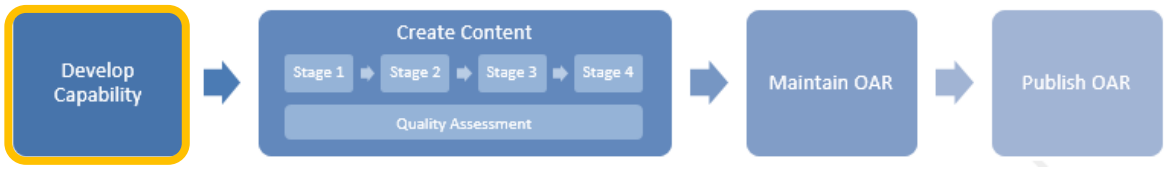
- **Stratified data samples**

  Different geographical environments present different challenges when populating addresses. Variation would be expected in the completion rate, assessment time and quality between different areas. Therefore, it was vital to ensure that a good stratification of these environments were assessed so that conclusions could be accurately made.

- **Quality Assurance**

  The desk based and field teams validated addresses completed by the automation stage within the immediate vicinity of the record they were manually geo-processing. For example addresses on the same street or in the same building. This provided additional assurance of the creation process and data quality analysis activity.

3. DEVELOPMENT OF OAR CAPABILITY



The necessary capability to create, maintain and publish a new OAR would need to be established before any content creation could begin. Where possible, existing processes and capabilities would be implemented, however to ensure that the solution is RM IP free some changes would be required. The Pilot explored what capability development would be required and the costs associated with these activities. There are three core areas which would need development of capability:

- **Creation Capability:** development of tools and technical infrastructure required to build the solution *(described further below)*.
- **Ingestion Capability:** development of processes to remove existing data containing RM IP and replace with newly created RM IP free data *(described fully within section 5 – Maintenance of the OAR).*
- **Publication Capability**: to fulfil the proposed solution in the same form and formats as AddressBase is currently supplied, a mechanism for delivery would need to be created. If the OS publication platform was used, development of a whole new capability may not be required *(described fully within section 6 – Publication of the OAR).*

Creation Capability: Tools & Infrastructure

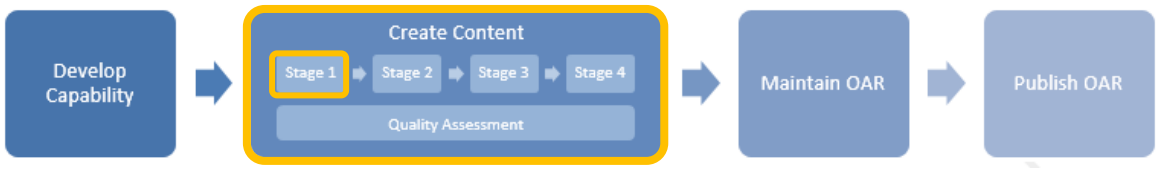
A production system would be required to collect and manage data during the initial creation of the OAR. It is anticipated that the content creation phase would take a number of years to complete therefore the production system would need to be robust and maintained. The technical infrastructure would have to meet the following requirements in order to be fit for purpose:

- The ability to manage a very large data set (over 32 million records)
- The ability to synchronise data from multiple data sources (automation, desk based geo-processing, field geo-processing) and ensure alignment between these data sources
- The ability to enable interrogation of the data to support GeoPlace with data maintenance and improvement activities including the investigation and resolution of queries, changes to the Specification or ad-hoc improvement programmes of work

It is estimated based on projects of a similar scope and magnitude that the design, development, testing and management of building the required infrastructure would be [redacted]. Annual hosting and maintenance would be [redacted] per annum.

4. CREATION OF OAR CONTENT

4.1. STAGE 1 - GENERATE CANDIDATE LIST



4.1.1 METHOD

A product specification was developed to enable an OAR product to be built. Please see Annex A for the full specification.

The method of creating the candidate list followed the steps below:

1. Identification of all live / approved address records[4] by using a flag inserted by Local Authorities.
2. Filtering of the above records undertaken by using a classification code inserted by local authorities, to ensure that the scope was limited to rateable objects and OWPAs.

4.1.2 RESULTS

The process identified **32,497,810** candidate records. At the interim point of the Pilot this was expected to be closer to 35 million records, however due to the refinement[5] of the OWPA definition it was found that many OWPAs within AddressBase did not fit the scope of the OAR and were removed from the candidate list.

4.1.3 COMPLETENESS

Using the above process, a high confidence level can be attributed to the likelihood that all candidates have been identified when comparing the selection to other Address products, namely the AddressBase family.

This does not mean that all rateable properties or objects are included due to different capture methods between Local Authorities and the Valuation Office Agency, but the above method should provide a list of equal completeness to that which could be obtained from the AddressBase suite. The following two examples demonstrate addresses of store rooms within single shops that are captured in VOA but would not be captured by Local Authorities of AddressBase: STORE CP8B, WEST QUAY SHOPPING CENTRE, SOUTHAMPTON or STORE LS4, WEST QUAY SHOPPING CENTRE, SOUTHAMPTON.

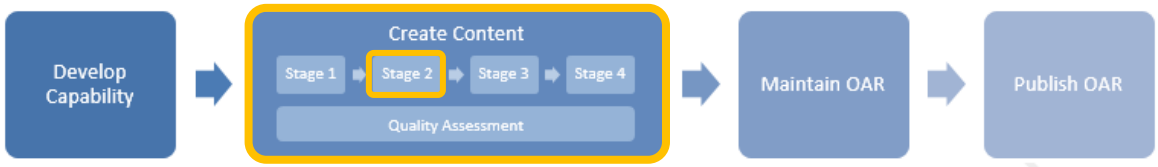
Within the desk based and field geo-processing stages of production there were circumstances in which addresses were identified that were not within the candidate list. For example, within an industrial estate a Local Authority may have only identified 1 commercial unit, but a surveyor might find more than this. The desk based and field teams captured additional addresses they encountered. Extrapolated to a GB scale, the estimate for unidentified addresses in the candidate list would be 9,000 records. It is believed that these addresses are within PAF but not AddressBase because of product specification issues.

4.1.4 TIME

With a stable specification and expert knowledge of the underpinning databases, the generation of the candidate list was a very quick process, taking less than one day. The process can easily be refreshed at any point in time, but may need updates if the specification were amended.

[4] Approved / Live records – These are all addresses which a Local Authority has marked as currently existing.
[5] In the Interim Report all OWPAs were included in the candidate count. A subset of these have been removed as they are not rateable e.g. ponds

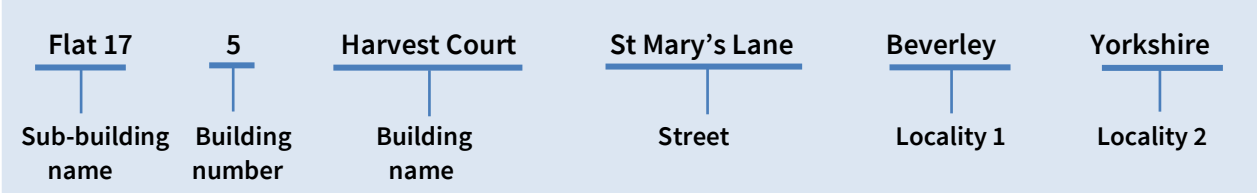
4.2 STAGE 2 - AUTOMATED GEO-PROCESSING



4.2.1 METHOD - AUTOMATED GEO-PROCESSING

Using the Specification (Annex A), the Automation Team focussed on populating each candidate record to the greatest extent possible. This included all fields from Building Name / Number to Local Names such as Town and Locality information. The algorithms developed were run nationally and due to the complexity in some cases have took up to 1 week to return results. The team only used OS data and did not include any third party datasets. Any OS data used was believed to be RM IP free.

A full OAR address would be made up of the following elements:



The following data sources have been analysed and interrogated to extract the required data for each of the address elements:

Sub Building Name / Number

1. **Local Authority only addresses with no RM equivalent**
   These records are addresses contained within the National Address Gazetteer (NAG) hub, but have no Royal Mail equivalent. This is due to different capture methods and requirements between the two data collectors. These records have therefore been used to help allocate sub building names and numbers.

Building Name / Number

1. **Use of OS core content**
   Using OS data sources, the automation team were able to assign building names and numbers to Building TOIDs[6]. This method was extended when compared to the interim report; and now included all address records for which OS have a single building name or number, whereas previously this only implemented one-to-one relationships between both the address and building name and number.
2. **Interpolation techniques to extrapolate a building number using other OS content where OS has not previously captured a building number**
   These techniques used OS MasterMap Topography Layer in several different ways and advanced GI techniques. Firstly, grouping together candidates of addresses, then using the already assigned building numbers from step 1, extrapolating additional building numbers assigning these to address candidates via the Building TOID.
   This technique was further advanced compared to the previous interim report with closed loop addresses (such as those around a Green) being included.

[6] TOID – This is a unique Identifier used by Ordnance Survey in many of its products. In the instance of this report the TOID referred to is the one assigned by Ordnance Survey to each and every building feature they have captured.

3. **Local Authority only addresses with no RM equivalent**
   This follows the same method as used for the Sub building number. These records are addresses contained within the NAG hub, but have no Royal Mail equivalent. This is due to different capture methods and requirements between the two data collectors. These records have therefore been used to help allocate building names and numbers. An example address would be: *The Flat, Dog and Duck Inn, Ladygate, Beverley, Yorkshire*.

Street

1. **Use of OS MasterMap Highways production system**

   Working in collaboration with the Department for Transport, OS has recently released a beta product to the PSMA and OSMA community, which combines the definitive attribute information contained within Local Government's National Street Gazetteer with the definitive spatial data from within OS MasterMap ITN product. This beta product was used to help assign street names to records within England and Wales, Scotland is currently not contained within this product. [redacted]

2. **Use of OS MasterMap ITN and spatial interpolation techniques**
   Where a street name could not be determined using OS MasterMap Highways production system method, ITN and the spatial function of 'Find My Nearest' was used to allocate a street name. The technique had two phases; firstly assigning a street name if all records on a given street returned the same result. Secondly assigning a street name to all records by taking the most prominent result when more than one street name was returned for records all on the same street as identified by the USRN.

Locality / Town / City

1. **Use of OS core content**
   Using core OS data higher geographies such as town and city information, were allocated to 15 million records. These were records which spatially fell within a named extent such as a town but did not fall into any other named area and therefore had a one-to-one relationship.
2. **Applying a hierarchy**
   As the above method only catered for a one-to-one relationship, a hierarchal system was then developed to cater for one to many relationships between a candidate record and a named extent. For example, Towns were assigned before Urban Developments. By applying this hierarchy multiple LOCAL _AREA_NAME values could be populated.
3. **Boundary Line**
   For any records which had not been assigned a LOCAL_AREA_NAME from the above two processes Boundary line was used to perform a spatial intersection. This meant that the remaining records could be allocated a higher geography attribute.

OWPA Records

1. **Manual Creation**
   OWPA records are not contained within Royal Mail data and therefore can be used RM IP free [redacted] Currently the address structure for OWPA records within AddressBase is made up of a text string description which includes the Street Name. Because this Pilot considered the street name to have the possibility of RM IP content, the text string for OWPA records was re-created. This means OWPA records only describe their use and not location e.g. 'TENNIS COURT' rather than 'TENNIS COURT 3M SOUTH OF BANKS ROAD'.

Third Party Data

The use of third party data was explored, but not implemented in any part of the methodology [redacted] Please see annex B.

[redacted]

4.2.2 RESULTS - AUTOMATED GEO-PROCESSING

4.2.2.1 COMPLETENESS

The algorithms and queries developed in the automation stage were run for Great Britain.

The team were successful in populating 21,611,044 (66.5%) records with each required primary field completed.

A further 10,447,642 million records (32.1%) had 2 primary fields completed and 441,282 records (1.4%) had 1 primary field populated. Figure 1 demonstrates the progress made on populating addresses for each of the candidate records.

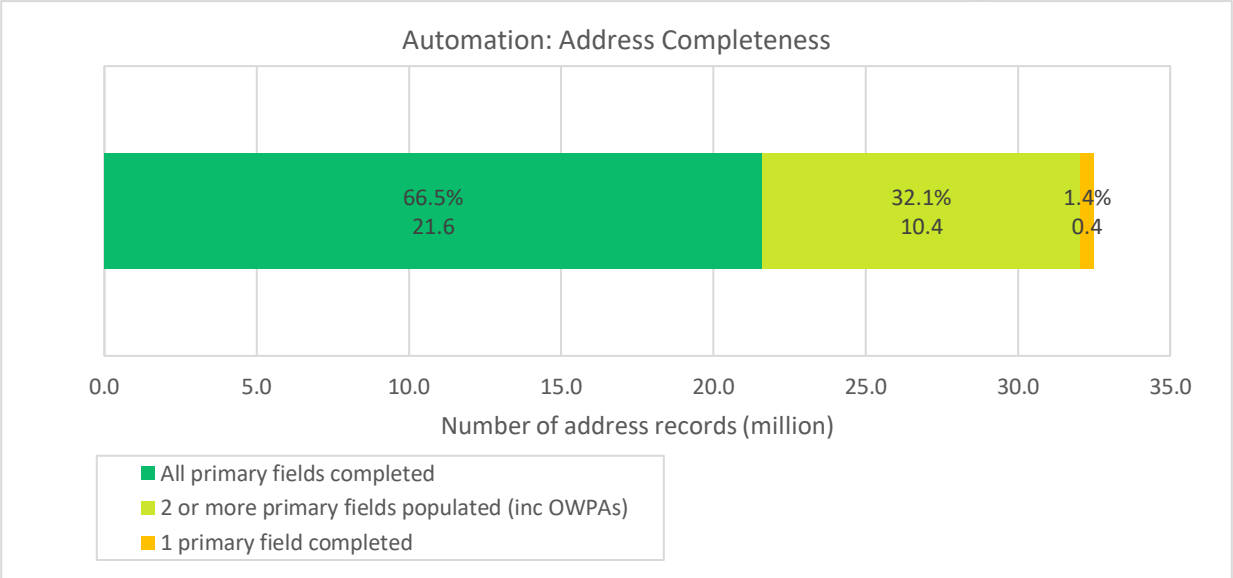


Figure 1: Breakdown of results from the automation stage

Addresses with incomplete primary fields (10.8 million records) were referred to the desk based geo-processing team for further assessment (see section 4.3).

Analysis of the automation output has shown that the levels of completed records vary between country, with England achieving the highest rate (72%), followed by Scotland (53%) and then Wales (29%). The majority of properties are within England (86%), therefore the overall completeness levels for GB have remained reasonably high at 66.5%. The reasons for the difference between countries include data availability (for example there is no coverage for Scotland in OS MasterMap Highways) and differences in types of geography within these countries. These findings indicate that more records for Scotland and Wales would be referred to the later stages of production.

The automation stage was most successful at achieving a complete address when capturing one-to-one relationships between OS cartographic text, building polygons and a single seed address. For example, in Figure 2 the OS cartographic text indicates that the building polygon highlighted in green contains 1 address seed and the property number is 50.

The automated stage was also successfully at extrapolating simple building numbers where there was no OS cartographic text within a building polygon, but cartographic text in neighbouring properties. For example, the property highlighted in blue on Figure 2 would be number 46.

Figure 2: Properties where automation was most successful

The following examples demonstrate types of instances where automation was unable to identify an address:

No Building Name / number for shopping centres

In Figure 3 (Westfield Shopping Centre) the only address reference discernible from OS data was the street name and overall building name of Westfield Shopping Centre. These elements could be used by automation but the example contains over 100 candidate addresses and therefore requires further work to identify the individual addresses within the location.

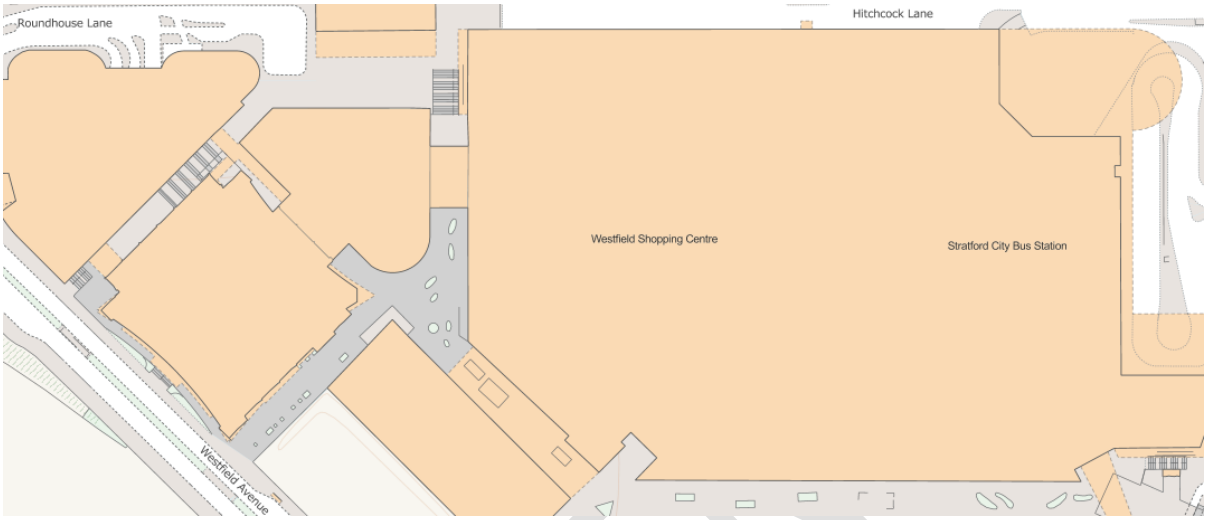


Figure 3: Westfield Shopping Centre

Airports and other large sites

OS content normally only captures the high level address for site locations for cartographic reasons.

In Figure 4 the airport buildings at Bournemouth Airport are only identified by text such as Air Passenger Terminal which is not granular enough for an automated system to create all of the subdivisions such as shops.

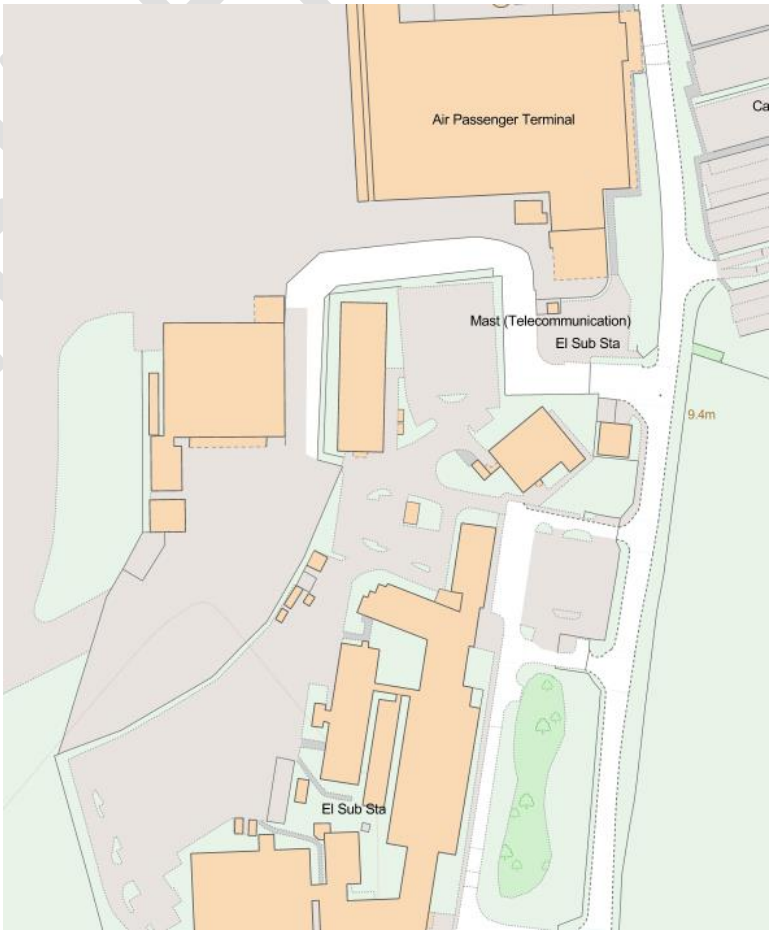


Figure 4: Bournemouth Airport

4.2.2.2 ACCURACY

To determine the accuracy of the completed automation results, records with all primary fields completed (21.6 million) were compared with AddressBase. The address text string was matched using the UPRN and any difference in characters recorded. For an address to be a valid "match" the following criteria had to be met:

- The building number had to be exactly the same between AddressBase and OAR and;
- The rest of the address had to be less than 6 characters different.

Where the address was found to be more than 5 characters different, desk based analysis was conducted to assess whether the address was still valid.

The overall accuracy of completed addresses that were found to valid was 92%.

The breakdown of this result is displayed in Table 1. It was found that 84% of records matched AddressBase. Of the remaining 16% that did not match AddressBase, 8% were found to be valid addresses and 8% were incorrect.

Table 1: Automation accuracy results

Address match between automation output & AddressBase	84%	16.3 million records
Address different to AddressBase but still valid	8%	3.2 million records
Address different to AddressBase and found to be incorrect	8%	2.1 million records

Through desk based investigation of a sample of the automation output, the causes of errors were analysed and a number of reasons identified why the automated process may have produced a different address to AddressBase. The most common example of addresses being different yet valid in both instances was where there were minor spelling and grammar differences in the street names. Another common difference identified was the locality. Through initial analysis it was noted that addresses within the OAR contained a more granular description of locality than AddressBase. For example:

OAR: 15a Albemare Road, South Bank, York vs AddressBase: 15a Albemare Road, York

The level of detail within the locality was deemed important for the OAR as the address would not contain a postcode. The result on accuracy however meant that the address strings between AddressBase and OAR did not match, but the addresses were still valid.

Examples of where an address is different from AddressBase and still valid:

Abbreviations, spelling & grammar

Minor differences such as changes in grammar, spelling mistakes in AddressBase and the lengthening of abbreviations, account for the majority of instances where AddressBase and OAR records appear different. In these circumstances, the addresses were often the same but the character difference was greater than 5. An example of this is as follows:

The OAR record was captured as: "23 Saint Oswald's Close, Kettering"

The AddressBase record is: "23 St. Oswalds Close, Kettering"

Both may be considered correct, but are different.

Road names

A common example is where the OAR has captured the road and AddressBase hasn't. Figure 5 displays a military base. The AddressBase record does not contain a road name, however the OAR does:

OAR: 3 Roberts of Kandahar Road, Oxendene, Warminster

AddressBase: 3, Oxendene, Warminster

In this instance the addresses are different but the OAR address is still valid.

Figure 5: Oxendene

OWPAs

As noted in the Method section, OWPA records have been created differently to avoid RM IP being introduced. Although the address is different it would be deemed that both addresses are correct. For example, in Figure 6 AddressBase gives a text string of - TENNIS COURT 13M FROM 25A VALLADALE. 35M FROM VALLADALE

Whereas the OAR will simply have TENNIS COURTS and then declare the correct Street and Local Area Name.

Figure 6: Example of an OWPA

For the 8% of incorrectly populated records, analysis has shown that there are several different circumstances where errors commonly occur. The following are examples of where the OAR automation results are incorrect:

Automated assignment of the wrong ITN TOID

Deriving a street name using ITN TOID data from AddressBase can cause an error in the resulting address. In Figure 7, addresses on St Bernard's Crescent have incorrectly been allocated an ITN TOID with the road name Dean Street. This is because the address seed placement (blue dot) is close to the back of the property. The result is that the address is populated with the nearest street, which is incorrectly Dean Street.

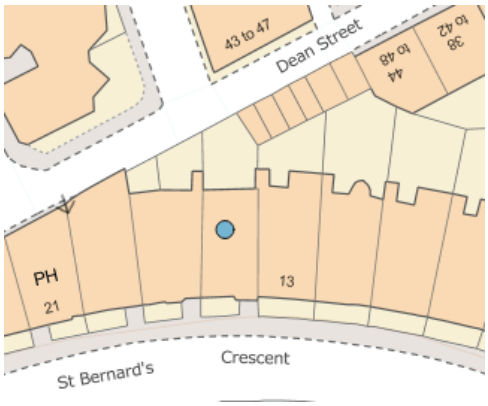


Figure 7: St Bernard's Avenue

Industrial estates

The automated process was poor at completing industrial estates. It has no way of capturing the name of the industrial estate in the address and it also does not capture any information on Organisations or the fact that the addresses are 'units,' thus at a first glance they appear as normal residential addresses.

Private roads

The addresses highlighted in Figure 8 are serviced by a private, unsurfaced track. No information for it is captured in our street gazetteers or datasets, as such they have been incorrectly assigned the main road (to the north) as their street. Not only is this incorrect but it also results in duplications.

OAR: 2 Manchester Road, Linthwaith, Kiklees

AddressBase: 2 Spring Road, Linthwaith, Huddersfield, HD7 5LT

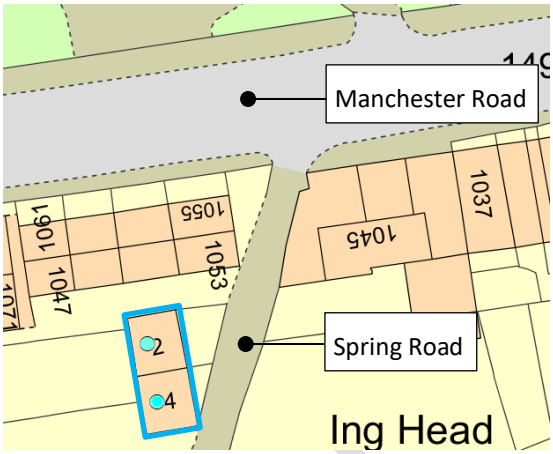


Figure 8: Ing Head

Terrace numbering

There have been several examples of multiple terraces with the same numbering sequence on the same road. For AddressBase records, the name of the terrace is given (shown in Figure 9) and the street ignored. In OAR records, the street is used and the name of the terrace is ignored. Resulting in multiple duplicates on a street with no distinguishing features.

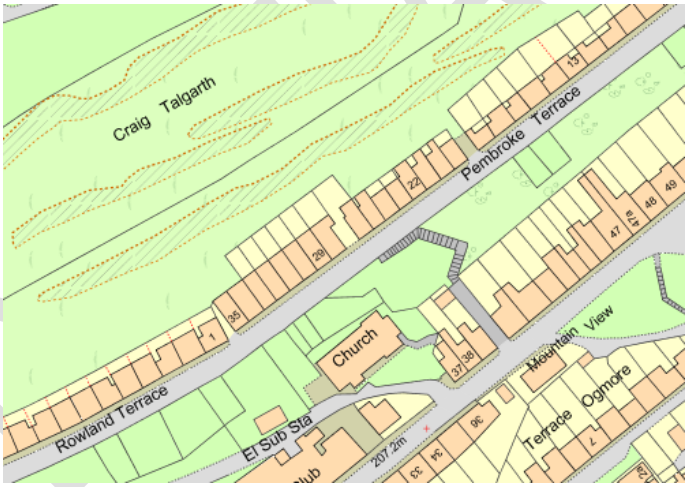


Figure 9: Terrace numbering

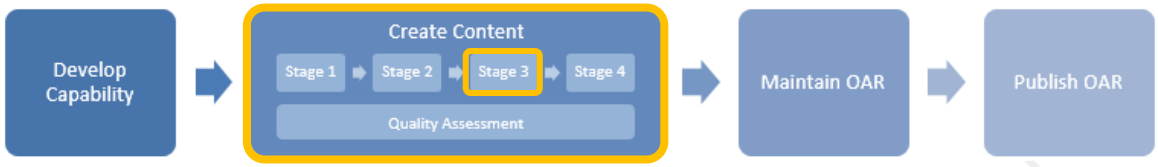
Through further development it may be possible for the automation team to refine their algorithms to exclude circumstances where an error is likely to occur. This has been possible during the Pilot for industrial estates, but would be more challenging and may not be possible for other examples.

It has been assumed that in a live production system, AddressBase would not be permitted for use of identifying potentially incorrect records. Therefore any incorrect completed addresses would be part of the populated OAR product. The full impact of this on overall OAR accuracy is described within section 4.5 Quality Assessment Summary.

4.2.3 TIME - AUTOMATED GEO-PROCESSING

The development of the automation algorithms and queries could potentially be an ongoing task in pursuit of high completion and accuracy results. In order to achieve valuable improvements, without costing more than manual intervention, the automation team estimate that 6 months of development would be required.

4.3 STAGE 3 - DESK BASED GEO-PROCESSING



4.3.1 METHOD – DESK BASED GEO-PROCESSING

Desk based intervention is required in scenarios where automated processing cannot identify addresses. The most cost effective method of manual intervention is desk based geo-processing where OS Editors make a visual interpretation of an address based on information available to them. In the context of a live production environment to create an OAR, the desk based editing team would be working on the 10.8 million records that could not be resolved by the automation stage.

The infrastructure was established to enable content to flow directly from the automation databases to the teams conducting the desk based geo-processing. So that the team could complete a significant sample of data within the timeframe of the Pilot, work began on desk based editing in parallel with the automation development. The automation element of the Pilot found this to be beneficial as lessons were fed back and processes improved, particularly in circumstances where an address had been incorrectly populated. However, in a live production scenario this would not be a practical way of working. The production stages would need to work in a waterfall process, one after another.

The OS Editors were tasked with populating candidate addresses with the following elements of an address:

- Sub building name and/ or number
- Building name and/ or number
- Street

Other address fields outside of those listed above were either already populated through the automation stage or were not primary fields. To identify the address elements, editors interrogated OS Imagery, OS Maps API service and Road Link Layer (Containing USRN attribution sourced from Highways). Figure 3 provides visualisation of the three data sources and how they can be used to identify an address, or a part of an address.

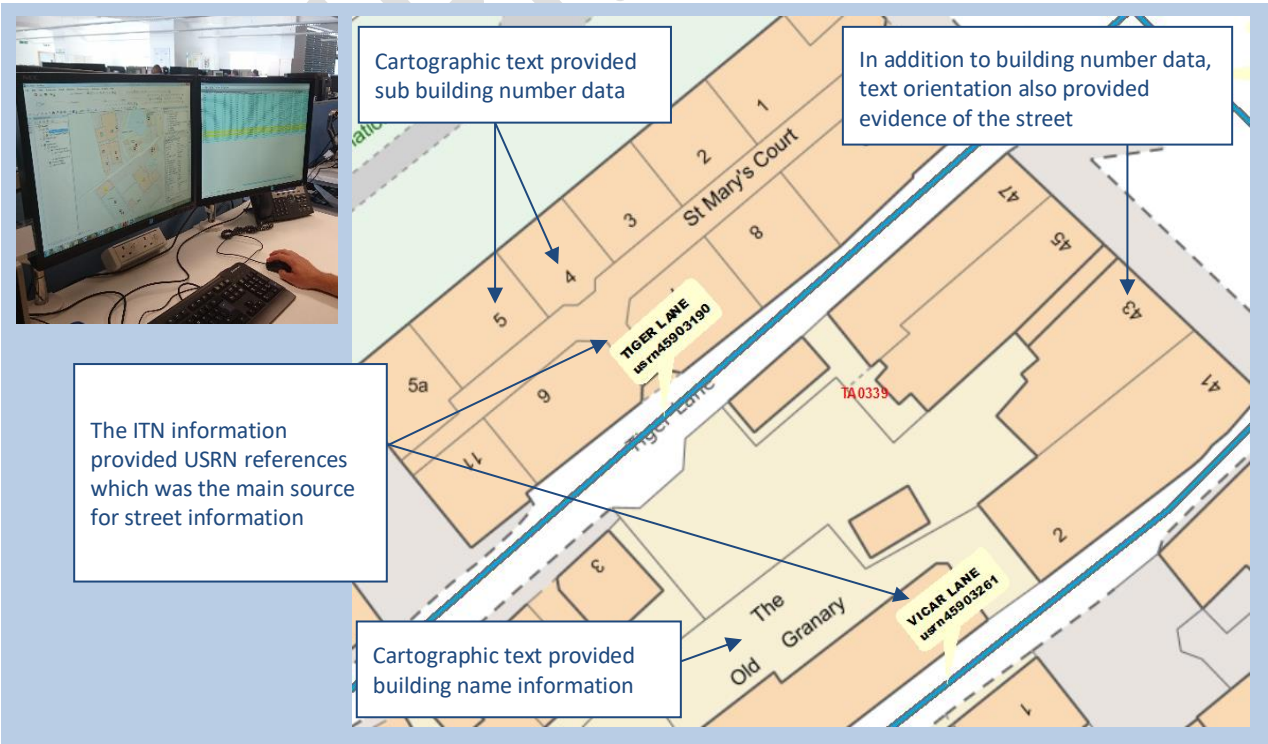


Figure 10: Representation of the data sources used in desk based geo-processing

When identifying a location to focus the desk based production effort, it was important to get a good stratification of urban, rural, residential, and industrial environments. These each presented different challenges to address capture and subsequently we expected to see variations in quality across them. To identify these areas, the 2011 Rural-Output Classification for Lower Layer Super Output Areas was used. This dataset was compiled for the 2011 census by the Government Statistic service in conjunction with Defra and OS. These areas were utilised as they provide a stratified break down of geographies, from dense urban to sparse rural and have clear definitions for each Geography type. More detailed information can be found here:
https://www.gov.uk/government/collections/rural-urban-definition

For this Pilot, the desk based effort focussed on the north east region of England. The areas displayed within Figure 11 were assessed.

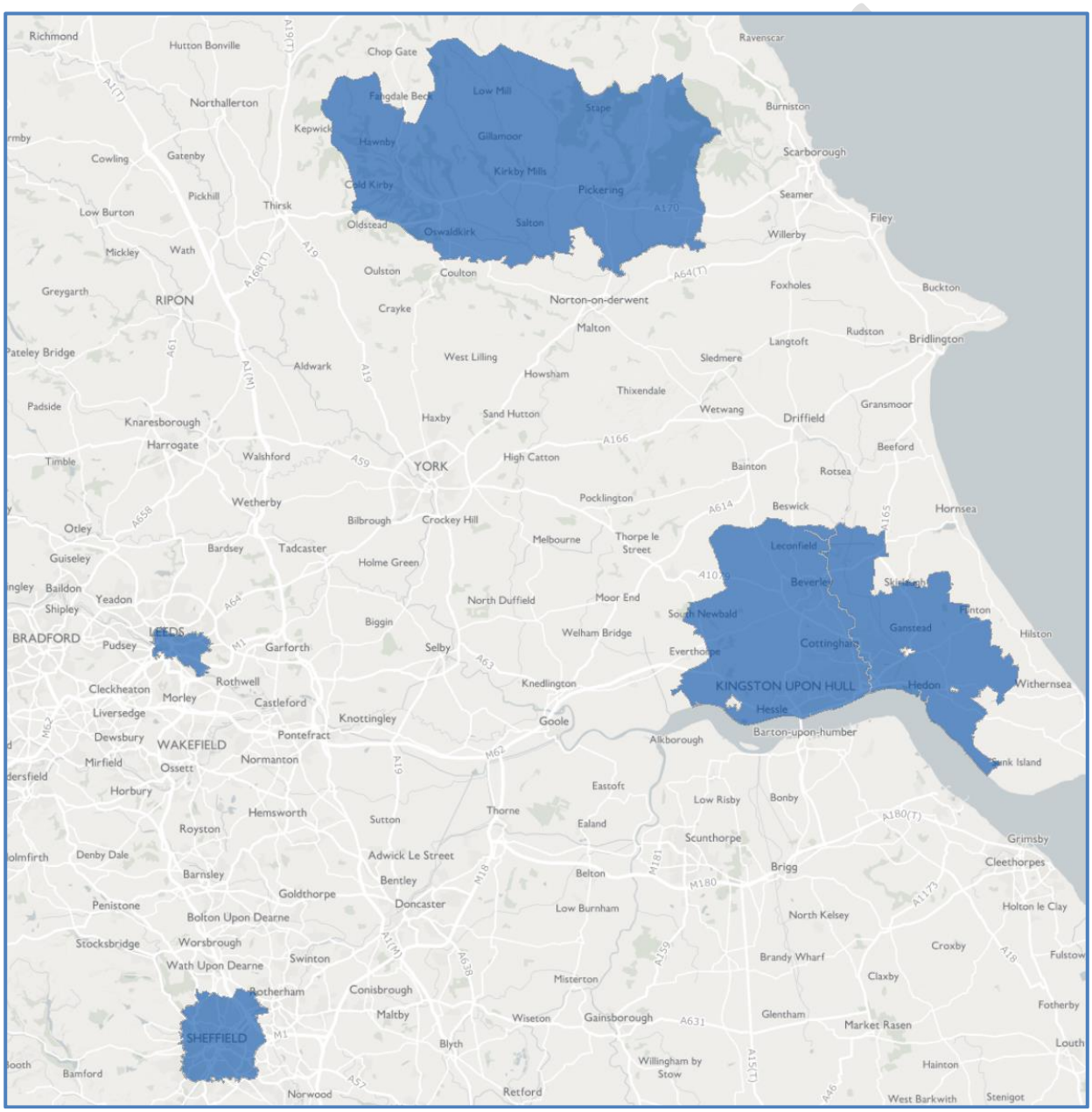


Figure 11: Geographic areas targeting by the desk based geo-processing team

4.3.2 RESULTS - DESK BASED GEO-PROCESSING

4.3.2.1 COMPLETENESS

Within Pilot timeframe the desk based team assessed a sample of 120,600 records.

Using the OS data sources available to them, editors were able to fully populate 91,800 (76%) addresses.

Editors were able to populate some but not all required address fields for 12,066 (10%) records and no additional information for 16,740 (14%) records. In both instances the records were referred to the field geo-processing team for further assessment. Figure 12 demonstrates the breakdown of results from the desk based geo-processing stage.

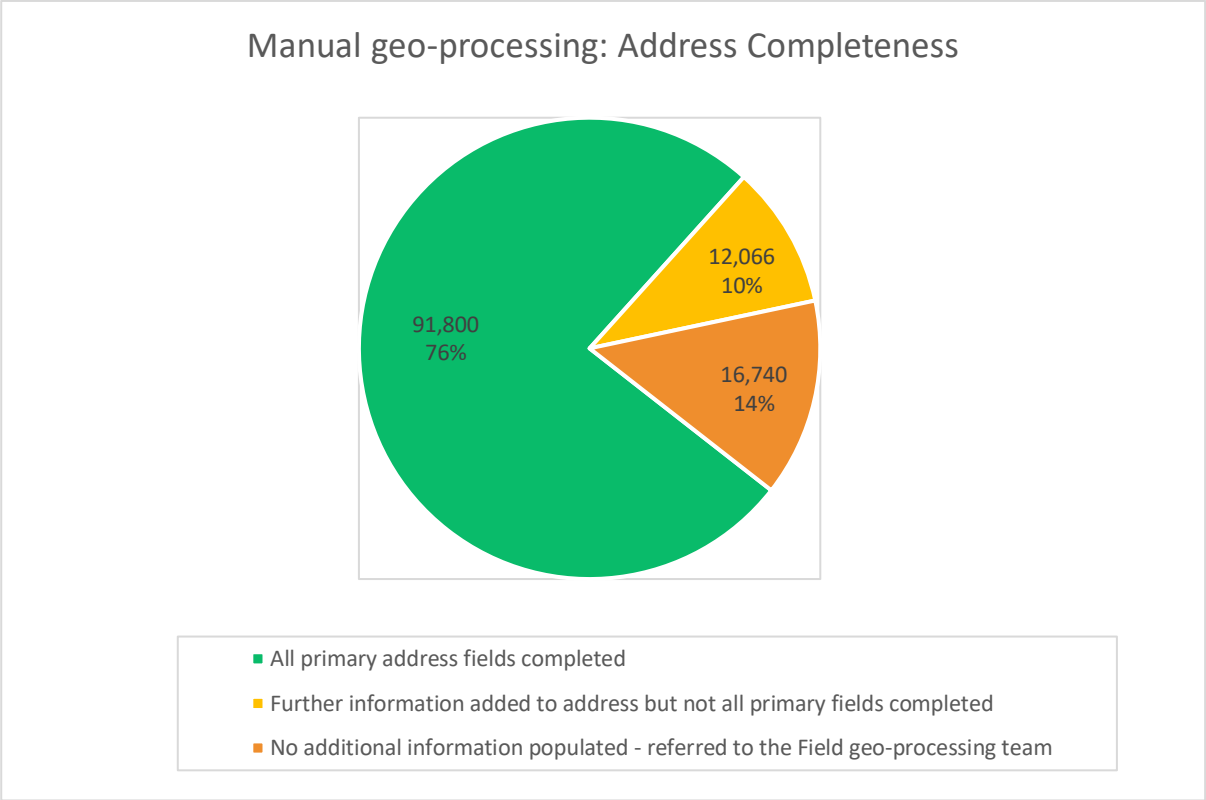


Figure 12: Breakdown of results from the Desk based Geo-processing stage

In addition to the addresses assessed above, the desk based team validated 21,750 records completed by the automation stage. The editors were able to validate these records as they were in close proximity of the addresses they were assessing.

Typically, the desk based team found residential terraces, urban areas and residential multiple-to-one addresses the most straight forward to complete with a high degree of confidence and speed. For residential terraces and residential multiple-to-one properties, a large number of addresses could be completed through bulk updates once the initial investigation of 1 or 2 addresses were completed. Due to the specification of the OS MasterMap Topographic data, urban areas have more features which can be referenced against.

The more challenging address types included the following:

- Instances where there is limited OS data available for sub-buildings and numbering
- Mixed classification with multiple-to-one addresses such as urban high streets where there is a mixture of commercial and residential properties within the same building
- Large sites including universities, hospitals, shopping centres, transport stations and industrial estates
- Rural areas where there are less features on which to reference

The following examples demonstrate instances where the desk based team were unable to complete an address:

Houses with names and numbers

In Figure 13 there are a number of buildings in The Chandlers. Due to the number sequence displayed/not displayed the desk based team were unable to determine what would be the correct addresses in the location.

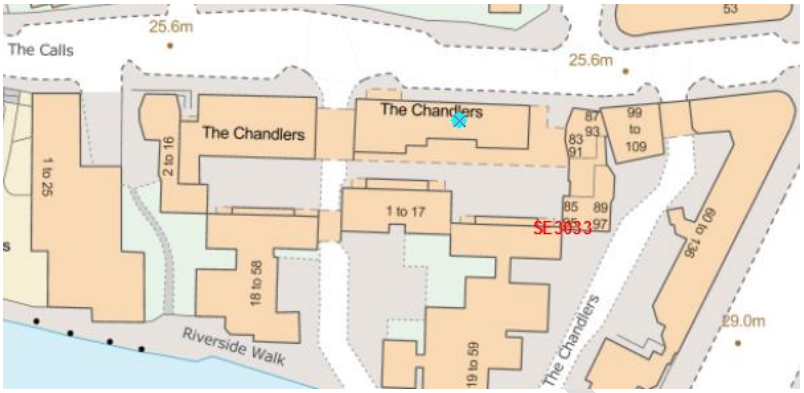


Figure 13: The Chandlers

Large sites

In Figure 14, the Howard Building at Sheffield Hallam University demonstrates where we have multiple candidate address locations (green boxes with cyan dot), but no sub-building information or numbers. The OS Topographic data does not provide the level of detail required to populate the full address. Therefore the addresses cannot be completed and would be referred to the field stage for a ground visit. This issue is common among large sites such as hospitals, shopping centres, transportation stations and industrial estates.

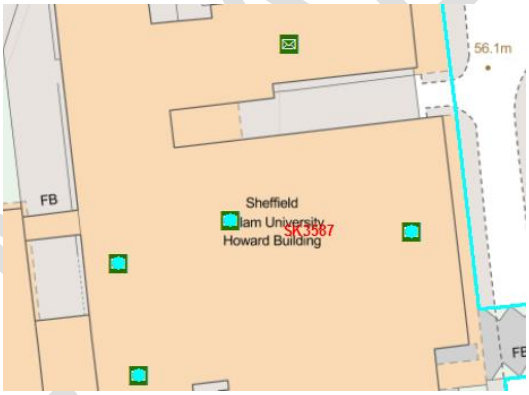


Figure 14: Sheffield University

4.3.2.2 ACCURACY

To determine the accuracy of the completed desk based results, records with all primary fields completed were compared with AddressBase. The address text string was matched using the UPRN and any difference in characters recorded. For an address to be a valid "match" the following criteria had to be met:

- The building number had to be exactly the same between AddressBase and OAR and;
- The rest of the address had to be less than 6 characters different.

Where the address was found to be more than 5 characters different, desk based analysis was conducted to assess whether the address was still valid.

The overall accuracy of completed addresses was found to be 89%.

Table 2 displays the comparison between the desk based results and AddressBase. It was found that 79% of records matched AddressBase. Of the remaining records 10% were found to be valid addresses and the remaining 11% were incorrect.

Table 2: Automation accuracy results

Address match between automation output & AddressBase	79%	72,501 records
Address different to AddressBase but still valid	10%	9,117 records
Address different to AddressBase and found to be incorrect	11%	10,095 records

The following examples demonstrate instances where the OAR address was different to AddressBase, but still valid:

Capture of the house name

In Figure 15, the desk based team have populated the property number, street name and locality. In AddressBase the house name is also captured therefore the addresses do not match. But both are still correct.

OAR: 5, Mill Rise, Driffield

AddressBase: 5, "Cayman", Mill Rise, Driffield

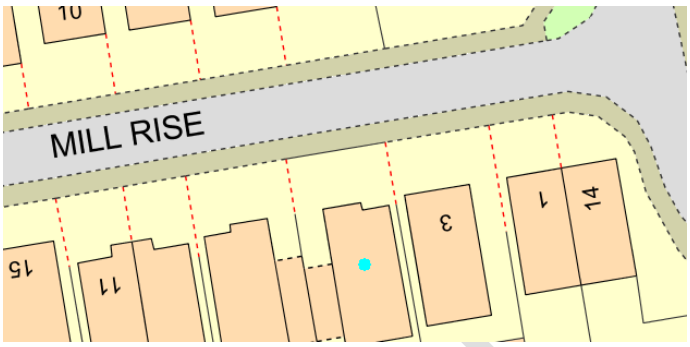


Figure 15: Mill Rise

Naming of flats

There are a variety of ways an address can be described and still be correct. In particular the use of the words flats or apartments. In this example the OAR describes the address as "Flat 1" and AddressBase describes it as "First Floor Flat". Both are correct.

OAR: Flat 1, 98, Park Grove, Hull

AddressBase: First Floor Flat, Park Grove, Hull

The following examples describe instances where the OAR is different from AddressBase and incorrect:

Incorrect street naming resulting in duplicate addresses

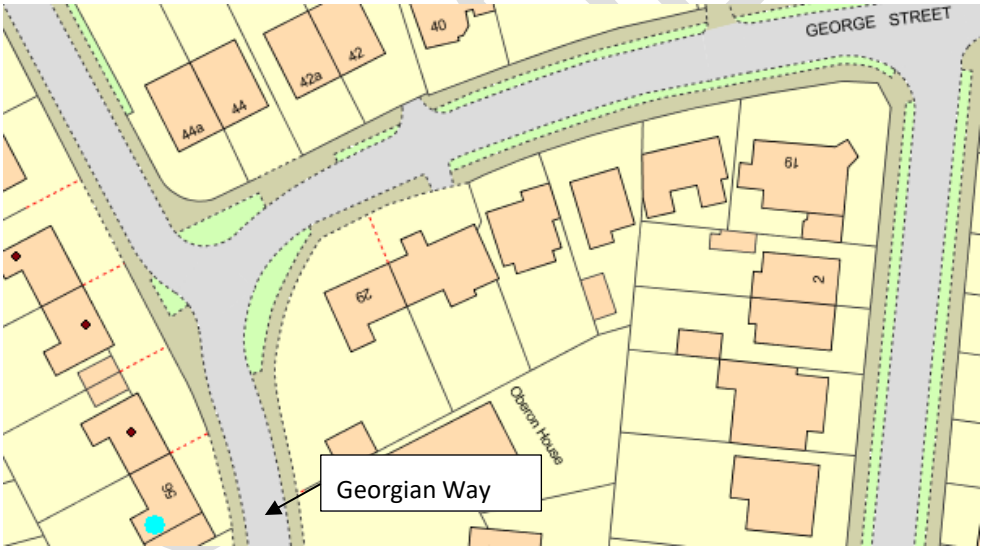


Figure 16: George Street

In Figure 16 the desk based team have populated the street name of property indicated by the cyan dot as "Georgian Way", as it appears that is the street the property sits on. The property is actually part of the numbering sequence from George Street and therefore should have the street name George Street. The real 56 Georgian Way is located some distance down the street therefore there would be two properties with the same address in the OAR.

OAR: 56, Georgian Way, Bridlington

AddressBase: 56, George Street, Bridlington

4.3.3 TIME - DESK BASED GEO-PROCESSING

The rate at which the desk based geo-processing team were able to assess records varied depending on the geography and type of address. As described under section 4.3.2, some address types were easier to complete and therefore faster to assess than others. In instances where it was not possible to populate an address, editors were able to quickly establish this and could refer to the field stage without expending too much time. Taking into account differences in address geography and type, editors were able to assess **500 records per day**.

For editors to assess and complete address records successfully to a high degree of accuracy, training and quality assurance monitoring would be required.

Training

Each editor would be required to complete a training package before beginning any production work. It was found in the Pilot that users who were less experienced with address data made significant errors resulting in re-work. These errors have been excluded from the sample results.

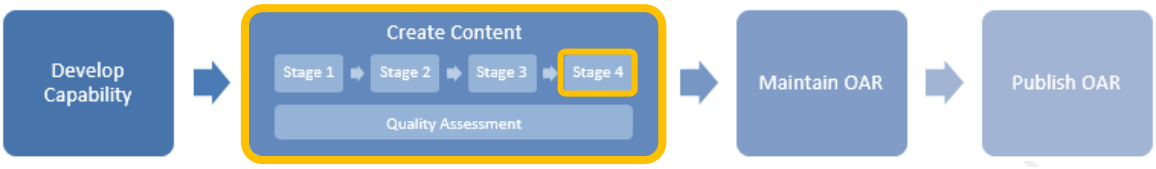
In light of the experiences and lessons learnt from the Pilot, it is estimated that editors would require 10 days initial training with best practice re-fresher short courses every 2 – 3 months. Editors would also need to successfully pass quality accreditation before editing live addresses.

Quality Accreditation

To ensure that editors are consistently demonstrating the required quality for the OAR, accreditation checks would be implemented. Editors would need to achieve the highest set level of accreditation before editing live data. All live data would continue to be quality monitored. Typically OS quality checks 10% of BAU activity.

The total time required to complete a national dataset is described within section 4.5 Quality Assessment Summary.

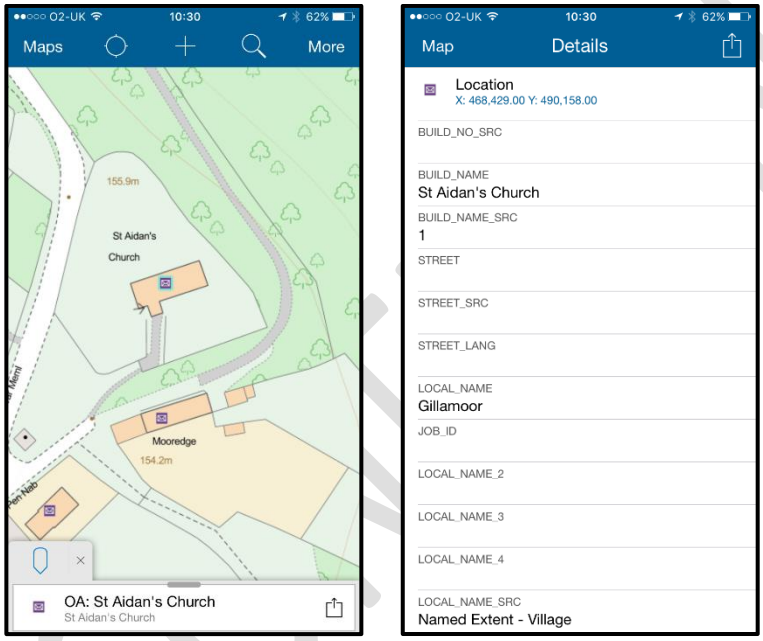
4.4 STAGE 4 - FIELD GEO-PROCESSING



4.4.1 METHOD - FIELD GEO-PROCESSING

In circumstances where the automation processes and desk based editing cannot resolve an address, it may be necessary for ground verification through a field visit. Our field resources were able to access data provided by the desk based geo-processing team. They conducted ground visits to properties and populated missing address fields where possible.

The geography of an area (rural or urban) had an impact on the data collection methodology implemented by the team. In urban areas, surveyors used the ESRI Collector App accessed through their smartphone to review addresses. The app directed them to a candidate point where the surveyor could make an assessment and enter any missing address fields. Figure 17 provides examples of the views surveyors would access.



The purple boxes on the map view are candidate locations which require an address to be populated. Surveyors were able to click on the box and complete the address elements for the record.

The second view shows the fields relating to the address.

Figure 17: Demonstration of the ESRI Collector App views

The process would take place online therefore any data entered would feed directly into the main production database. This was only possible when connected to a 3G, 4G or WIFI network and therefore was not an appropriate method for use in rural areas due to poor network coverage.

In rural areas surveyors worked offline using Toughbooks[8]. Data required would be downloaded in advance, and edits made offline whilst on location would be submitted to the production database at the end of the day. This method required careful planning to ensure simultaneous edits did not occur (two people editing the same area) and that enough work was allocated to prevent time wastage.

If this process were to be productionised alternative equipment might be explored. It was found that the mobile phone screen was often too small for context and the Toughbooks too big to carry for sustained periods of time.

In the Pilot OS Surveyors were used due to their experience in the field with address related activity.

[8] A Toughbook is a mobile computer built to withstand hazards associated with working outdoors such as weather or rough handling. OS Surveyors are all equipped with a Toughbook.

4.4.2 RESULTS - FIELD GEO-PROCESSING

4.4.2.1 COMPLETENESS

The Field team assessed 7,800 records. A sample of these records (2,240) were not assessed by the desk based geo-processing stage therefore have been discounted from any completion, accuracy or speed analysis. The excluded sample was made up of rural properties, and were assessed to test the rural methodology.

Using the OS data sources available to them, surveyors were able to fully populate 2,850 addresses (51%).

Surveyors were unable to identify any further address information for 2,700 addresses (49%). Figure 18 demonstrates the breakdown of results from the field geo-processing stage.

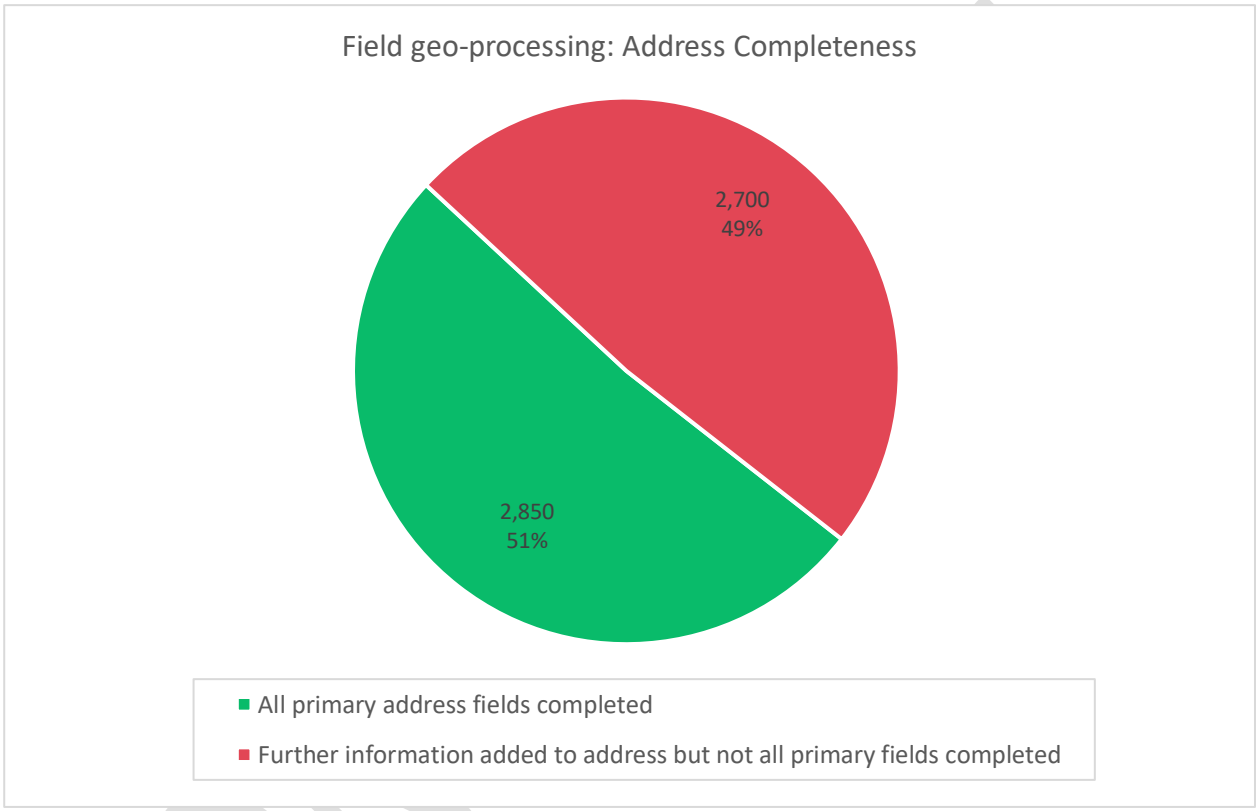


Figure 18: Breakdown of results from the Field geo-processing stage

The focus of the field effort was in urban / semi urban areas. According to analysis of the candidate records, approximately 80% of properties are located within urban areas therefore the sample represents the majority of address properties.

The completion rate for the field stages was found to be substantially lower than the desk based geo-processing stage. Generally, the records remaining after the automation and desk based stages were the most challenging to assess and complete. The following examples demonstrate commonly found circumstances where the field team were unable to resolve an address.

Lack of addressing information

A common scenario identified by the field team was where there were no distinguishing features on a property or surrounding properties to identify a house number. The example in Figure 19 demonstrates a property where there is no house number or name, and the occupant of the property was not available to answer any questions. The surrounding properties on the street only had house names. Therefore it was not possible to complete the address.

Figure 19: House with no number

Stacked addresses

The field team could confidently resolve the address points which were correctly positioned throughout the Trinity shopping centre, assigning them the street from which the shop was accessed.

However, in the middle of the shopping centre there were a large number of addresses stacked on a single point (identified by the arrow in Figure 20). They were not able to fully resolve these as it is impossible to tell which shop/unit they each reference and therefore which road name should be assigned. This was a common occurrence for shopping centres, train stations and other large commercial areas.

Figure 20: Trinity Leeds

Street signs

Where an A road runs through a town centre, there is often no street sign for the road name. The street may be commonly known as "Main Street" but there is no sign to indicate this. Therefore, it would not be possible for the field team to complete the address.

OAR: Lilac Farm Cottage, Lewisham

AddressBase: Lilac Farm Cottage, Main Street, Lewisham

Some multiple-to-one properties

Figure 21 shows a single large building with three separate 'building names' and a corresponding range of flats for each. There was no way of telling which 'building name' any of the addresses located inside the complex belonged to. Furthermore there were a different number of candidate addresses than the ranges given in MasterMap suggested, likely due to the building having been renovated.

The field team could not fully resolve any of the address candidates in this instance.

Figure 21: multiple-to-one addresses

4.4.2.2 ACCURACY

To determine the accuracy of the completed field-based results, records with all primary fields completed were compared with AddressBase. The address text string was matched using the UPRN and any difference in characters recorded. For an address to be a valid "match" the following criteria had to be met:

- The building number had to be exactly the same between AddressBase and OAR and;
- The rest of the address had to be less than 6 characters different.

Where the address was found to be more than 5 characters different, desk based analysis was conducted to assess whether the address was still valid.

The overall accuracy of completed addresses was found to be 83%.

Table 3 displays the comparison between the field results and AddressBase. It was found that 36% of records matched AddressBase. Of the remaining records, 47% were found to be valid addresses and the remaining 17% were incorrect.

Table 3: Field accuracy results

Address match between automation output & AddressBase	36%	1,026 records
Address different to AddressBase but still valid	47%	1,336 records
Address different to AddressBase and found to be incorrect	17%	484 records

Analysis has shown that the main reason for inaccuracies was where OS surveyors inferred an address building number from a nearby property, and this has been incorrect. It may be possible to revise the field methodology to prevent surveyors from inferring numbers. This could however have a negative impact on the number of completed addresses.

4.4.3 TIME - FIELD GEO-PROCESSING

The rate at which the field geo-processing team were able to assess records varied greatly between urban and rural geographies. As would be expected in a rural geography the distance between addresses was often found to be quite large. In some instances, addresses were over a mile from where a surveyor could park and access could only be gained on foot. Within urban areas where candidate sites were within close proximity, assessing addresses was a quicker process. On average, including travel time, surveyors were able to assess **60 records per day**.

Training

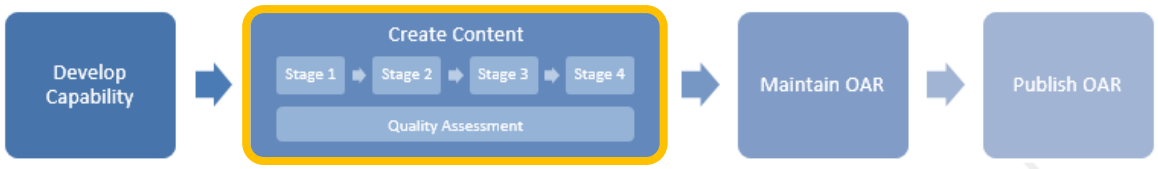
OS surveyors are experienced in working with address information and are located throughout England, Scotland & Wales. The nature of the addresses requiring population at this final stage of production are typically the more challenging address types and require experience to assess them successfully. In order to reduce travel costs and create an efficient way of working, all of the OS surveyors would need to complete an OAR training package. It is estimated that the initial training would be a 5 day course with regular best practice refreshers. For newly recruited staff the training period would be longer.

Quality Accreditation

To ensure that surveyors are consistently demonstrating the required quality for the OAR, accreditation checks would be implemented. Surveyors would need to achieve the highest set level of accreditation before editing live data. All live data would continue to be quality monitored. Typically OS quality checks 10% of BAU activity.

The total time required to complete a national dataset is described within section 5 - Content Creation Results Summary.

5. CONTENT CREATION RESULTS SUMMARY



The results of each stage of the content creation phase are summarised in Table 4, including the completeness and accuracy findings from the quality analysis. The automated result provides a figure against the entire candidate dataset whereas both the manual and field elements are based on a sample of that dataset that could be investigated in the given timescale.

Table 4: Pilot results summary

Content creation stage	Number of records assessed	**% of completed addresses**	No completed addresses	**% accuracy**	No of **incorrect** addresses	Geographical extent
Automation	32.5 million	**66.5%**	21.6 million	**92%**	1.7 million	GB
Desk based	120,600	**76%**	91,800	**89%**	1,098	Sample
Field	5,550	**51%**	2,850	**83%**	484	Sample

In reviewing the data within this table there are some points to note:

- Field geo-processing appears to be able to complete an address in half of the data this process looked at. As per examples provided earlier in this document against the field geo-processing stage, it was identified that the low completion % was mainly due to the types of records left after the previous production stages. Only the 'difficult' addresses remained.
- The spread of error is not consistent across the dataset. There are a few factors that skew the quality achieved as represented in the examples provided in this document:
    - Based on the analysis of the entire automated geo-processing output the automated process was most successful in urban and suburban areas. As the addresses became more rural the process became less capable of completing an address successfully. Less property numbering and lack of road naming are key factors. These issues were also identified in both the manual and field geo-processing results
    - The creation process was also less successful in dense urban/city centre geographies with less 'one-to-one' address relationships, and more multiple-to-one addresses (both residential and commercial geographies)
- For the reasons mentioned above (and due to lack of availability of a national streets dataset for Scotland) there is a regional skew with highest accuracy achieved in England then Wales with the lowest achieved in Scotland

T-o understand the impact of the completeness and accuracy of each production stage on a national scale, the numbers have been extrapolated. The content creation stages would happen sequentially therefore the incomplete address from the automation stage would feed into the desk based stage, and the incomplete addresses from the desk based stage would feed into the field stage.

As with any extrapolation exercise, there is a risk that the completeness and accuracy numbers we have calculated do not reflect the real world. As the automation stage was completed on a national scale we have a high level of confidence in the numbers, however for the desk based and field stages of production the risk is much greater. Every effort was made to cover a range of property types and geographies and complete a statistically significant number of addresses, however there may be address scenarios that our production teams did not encounter which could change the results. Figure 11 provide an extrapolated summary of each stage of production for GB.

Table 5: Extrapolation of results to national scale

Content creation stage	Number of record requiring assessment	**% of completed addresses**	Number of completed addresses	**% accuracy**	Number of **incorrect** addresses	Number of addresses requiring further processing
Automation	32.5 million	**66.5%**	21.6 million	**92%**	1.7 million	10.9 million
Manual	10.9 million	**76%**	8.3 million	**89%**	910K	2.6 million
Field	2.6 million	**51%**	1.3 million	**83%**	230K	**1.3 million**
TOTAL			**31.2 million**		**2.9 million**	**1.3 million**

Number of candidate addresses: **32.5 million**

Completeness in comparison to candidate list: **96% (1.3 million records incomplete)**

Estimated accuracy of completed OAR product content: **90.8% (2.9 million incorrect addresses)**

Therefore 4.2 million (13%) known records will either be incomplete or incorrect in the OAR.

The high number of addresses remaining incomplete following all processes (1.3m) forms a significant piece of work with no clear idea or plan of what could be done to tackle it. OS were only able to identify the high number of incorrect addresses (2.9m) by using AddressBase as a reference dataset for comparison only. It is assumed this would not be possible as part of a production process and therefore those records would be in error and be part of the product. We would not be able to identify them without incurring a RM IP violation. Both the incomplete and incorrect records would form a significant risk to the success of any product.

In the creation process we assumed that addresses existing at the same site would not require unique identification. This approach is suitable for a stand-alone creation process not one that would require maintenance and integration with other suppliers.

Assigning the correct UPRN to multiple-to-one properties

Multiple-to-one properties (where more than one address resides in one building feature) presented a major challenge to the OAR Pilot in terms of assigning the correct the Unique Property Reference Number (UPRN). Without cross-checking the AddressBase address, for all content creation stages there was no means of accurately assigning a UPRN to a created address.

The primary address information could predominantly be captured through automation and desk based activities, with some sub building information being allocated in this manner also (as per Figure 21). But in many instances field based resources were required to allocate sub building information.

Figure 22 demonstrates that automation and desk based techniques would be able to deduce the primary building name (Crown Mews) and that the six addresses contained within the block should be 1-6.

But automation, desk based or field activities would be unable to allocate the numbers 1-6 to the correct UPRNs.

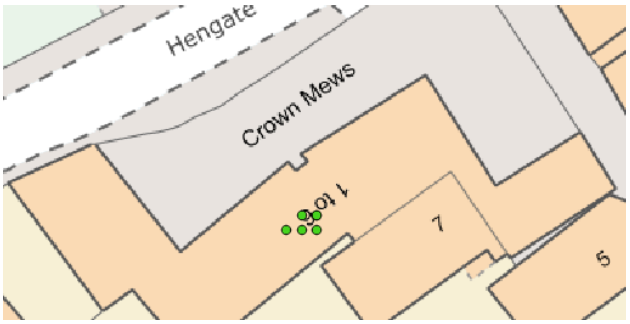


Figure 22: Crown Mews

The example in Figure 23 would enable automation and desk based activities to assign both primary and secondary address level information.

But there is a high likelihood that the address information although correct, could be assigned to the incorrect UPRN.

Figure 23: Goths Lane

In both Figures 22 and 23 the UPRNs have been assigned to an address by chance. There is a high likelihood that the address information although correct, could be assigned to the incorrect UPRN. The production teams followed a consistent methodology of assigning UPRNs by ordering the UPRNs and addresses sequentially. For example, the smallest UPRN was matched to Flat 1, second smallest Flat 2 etc. It was thought that some Local Authorities who assign the original UPRN may have followed a similar method, therefore there may be a higher probability of create a correct match, than just chance. Address custodians do not follow a consistent method of assigning UPRNs, therefore it is know that a proportion of UPRNs will be incorrect.

In order to understand the proportion of UPRNs likely to be incorrect, statistical analysis of the probability of assigning a UPRN was calculated. Please see Annex C for detail on how the mathematical probability was calculated.

It is estimated that the percentage of correct UPRNs in multiple-to-one buildings nationally would be 22.3%.

In summary, multiple-to-one address records represent 32% of OAR records. On average these addresses would be assigned to the correct UPRN in approximately 22.3% of cases. This percentage may rise in areas where Local Authorities have assigned UPRNs sequentially.

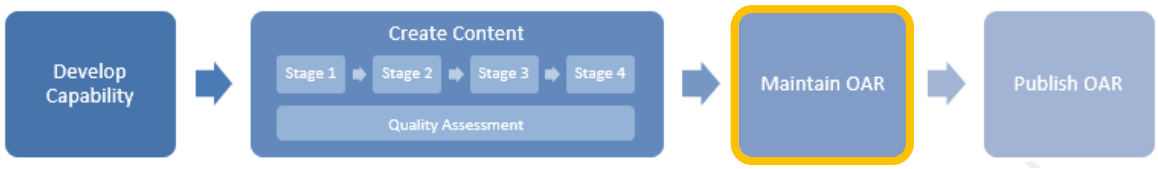
Manual validation of a small sample demonstrated that the UPRN could be assigned correctly in up to 54% of multiple-to-one properties, however it is unknown how applicable this number is GB wide. Further analysis and sampling would be required.

Potential content creation improvements

The findings summarised above are based on the results of this short Pilot. It may be possible to identify methodology improvements for each of the production stages including investigating methods of identifying incorrect records without using AddressBase or RM IP contaminated data. However, further time and investment would be required to do so and it is not known at this point how fruitful these investigations would be. OS experience with address content improvement projects have demonstrated that it can be challenging and time consuming to successfully identify errors and the methods required to correct them.

A potential completeness and accuracy improvement option could be to invest in OS's core products and content, for example improving the internal capture and representation of the properties that contain multiple-to-one addresses, e.g. shopping centres / train stations. This would overcome many of the data issues encountered in this project. Further work would be needed to determine the required investment and quality impact on the OAR data. This is currently out of scope for this Pilot.

6. MAINTAINANCE OF THE OAR



GeoPlace currently manage the maintenance of the National Address Gazetteer (NAG) which contains 40 million addresses. In considering the activities, costs and timescales required to manage the maintenance of a new OAR the following assumptions have been made by GeoPlace:

- The data would be maintained through existing NAG maintenance processes, with an extraction of new records (Royal Mail IP free) from Authority Updates.
- The existing NAG processes (data collection from LAs, inclusion of VOA, PAF etc.) and the creation of AddressBase and the NSG (and Streets Data for the Highways Product) would need to continue. Therefore GeoPlace would not expect to find any cost savings – only additional cost of managing the new OAR.

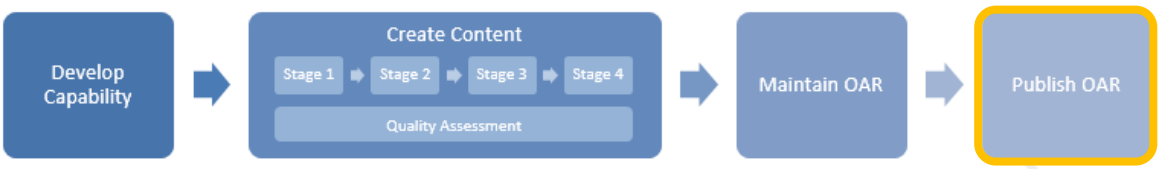
To provide an appropriate maintenance service for an OAR the following activities would be required:

- The design and development of data transfer formats, import, export scripts, reporting, auditing and automation
- The development of database tables
- Validation of the OS MasterMap data generated against NAG

It is important to note that in order to be able to extract new Royal Mail IP free addresses from Authority Updates provided by Local Authorities, authorities would be required to identify these new addresses primarily through the inclusion of street naming and numbering cross-references. The flagging of these records would denote the source as being entirely local authority IPR rather than having any contamination from 3rd party data. To implement this attribution would be a significant acceleration of a set of non-mandatory business processes GeoPlace have been promoting to authorities over the last 10 years. Currently GeoPlace receives Street Naming and Numbering cross references from approximately 70% of authorities but it is unknown whether these represent a complete picture of all Street Naming and Numbering records from these authorities. The remaining 30% of authorities would need to implement significant business changes to their address data maintenance processes to fully link Street Naming and Numbering with local gazetteer maintenance. This carries some significant cost burden which local government would expect to be covered under this proposal.

GeoPlace estimates that an initial cost of approximately £2.5m would be incurred by authorities in migrating the remaining 30% of authorities to new business processes and cementing these processes in those authorities who currently provide Street Naming and Numbering cross referencing. This could take 12 -24 months to implement fully.

7. PUBLICATION OF THE OAR



To fulfil customer needs the OAR could be provided to customers through a variety of publication options. Depending on what the publication requirement is, these will have differing costs and timescales to develop.

It is assumed that the publication requirement for the OAR would be to provide a publication offering which matches the current AddressBase fulfilment options.

AddressBase is currently supplied to customers every 6 weeks and is offered in a range of geographies including GB datasets, down to bespoke customer defined areas of interest for regional and local users. These geographies can be supplied as Geo chunks (tiles) or non Geo chunks (seamless). Customers have a choice of two formats (GML3 and CSV) and can access the data via Download, FTP, DVD or Hard Drive. There is a project in progress to develop the capability to deliver daily supply, likely to be completed early 2017.

To provide the services described above for the OAR, a number of small enhancements to the infrastructure would be required. These include development of components of the OS fulfilment systems to accommodate the OAR. Based on past estimates for projects with similar publication requirements, the appropriate costs and timescales have been provided for the publication capability development and publication annual maintenance within section 8 OAR Cost.

8. OAR COST & TIME SUMMARY

To create and maintain the solution described, the following cost estimates have been collated. These are direct production costs and do not include any overheads or margin. The additional charges would need to be considered using the government to government recommendations in "managing public money". All costs are based on the OS 2016/2017 pay rate. The costs do not take into account any required data improvement activities.

Table 6: Cost and timescale breakdown

DEVELOP CAPABILITY			
	Technical Infrastructure	[redacted]	12 months
	Maintenance capability	[redacted]	Up to 24 months
	Publication capability	[redacted]	6 months
CREATE CONTENT			
	Automated geo-processing	[redacted]	6 months
	Desk based geo-processing	[redacted]	24 months
	Field geo-processing	[redacted]	36 months
	TOTAL	[redacted]	**5 ½ years**
MAINTENANCE			
	Product maintenance	[redacted]	
	Infrastructure maintenance	[redacted]	
	Publication maintenance	[redacted]	
	TOTAL PER ANNUM	[redacted]	

These activities could be run in parallel

Field based activity could begin 6 months after desk based geo-processing, and from that point onward run in parallel

The total cost to develop capability and create OAR content would be [redacted] The OAR would have completeness and accuracy levels of those described within section 5 – Content Creation Results Summary.

The estimated cost to create an OAR has decreased from the original Solution Document estimate. This is due to several reasons:

- The project requirements are now more clearly defined and understood
- The use of additional OS datasets which increased the number of completed records through automation, therefore resulting in less records being referred to the costlier desk based and field stages of production.
- The desk based geo-processing team were able to process records significantly faster than anticipated. The change in speed was due to the type of records being assessed, in that they were relatively straight forward in comparison to the usual addressing improvement activities that editors are involved with. Editors were also able to make a very quick decision regarding whether they would be able to populate an address with the information available to them. Therefore, did not waste time on attempting to resolve an address that was not viable.

It is estimated that the time required to develop and create content for an OAR would take 5 and a half years. Some activities could run concurrently, for example capability development activities, however the automated geo-processing could only begin after the technical infrastructure had been completed. The automated geo-processing would also need to be finalised before the start of any manual geo-processing. Field activity could commence after 6 months of desk based geo-processing.

The time estimates have been based on OS resource feasibility. To deliver an OAR solution faster than the timeframes described within Table 6, significant increases in OS editors and field staff would be required. Table 7 demonstrates the number of staff required to complete the desk based and field geo-processing activities within 1, 2 and 3 years.

Table 7: Desk based and Field geo-processing resource estimates

	Number of FTE's required (including Quality Checkers)		
	1 year	2 years	3 years
Desk based geo-processing: Editors	120	60	40
Field geo-processing: Surveyors	240	120	80

Significant increases to staffing would have an impact on cost estimates in terms of recruitment, equipment and work space availability. There could also be a detrimental impact on product quality depending on the experience of newly recruited staff. It would be expected that training and quality checks would be increased for new staff to ensure quality levels did not decrease.

9. CONCLUSIONS

The Pilot has proved to be a useful exercise in understanding the potential to create an Open Address Register which contains no Royal Mail IP from OS content. To re-iterate, the Pilot was run on the assumptions stated in section 2.3, [redacted]

7.1. Completeness & accuracy

The results of the prototype, when extrapolated up to a national scale, are:

Content creation stage	Number of record requiring assessment	**% of completed addresses**	Number of completed addresses	**% accuracy**	Number of **incorrect** addresses	Number of addresses requiring further processing
Automation	32.5 million	**66.5%**	21.6 million	**92%**	1.7 million	10.9 million
Manual	10.9 million	**76%**	8.3 million	**89%**	910K	2.6 million
Field	2.6 million	**51%**	1.3 million	**83%**	230K	**1.3 million**
TOTAL			**31.2 million**		**2.9 million**	**1.3 million**

Number of candidate addresses: **32.5 million**

Completeness in comparison to candidate list: **96% (1.3 million records incomplete)**

Estimated accuracy of completed OAR product content: **90.8% (2.9 million incorrect addresses)**

Therefore 4.2 million (13%) known records will either be incomplete or incorrect in the OAR.

A number of trends have been identified with regards completeness and accuracy:

High levels are achieved where:

- There is one address within a single property (a one-to-one address), and the name / number has already been captured by OS
- A one-to-one address, and the number can be interpolated from nearby addresses where the number has already been captured by OS
- An address is an OWPA, e.g. a bill-board / electricity sub-station
- An address is unique to Local Authority and not captured by Royal Mail

Lower levels are achieved where:

- There is more than one address within a single property (multiple-to-one addresses), e.g. a block of flats / shopping centre
- The address is accessed from a private shared drive (e.g. a block of flats / industrial estate)
- The address only has a name
- The address is uniquely identified only by an organisation name

Further in depth analysis of the data has identified that if the OAR was limited to particular subsets of addresses (e.g. one-to-one addresses where the name / number had already been captured by OS in a non-address product), whilst completeness would be reduced, the accuracy of these records would increase.

7.2. Costs

From the Pilot it has been estimated that the above results could be achieved in 5 ½ years, at a cost of [redacted] with ongoing maintenance fees of [redacted] This is lower than the original anticipated costs described within the solution document due to:

1) Improved efficiency in the creation methodology

2) Use of existing address data that contains no RM IP

Whilst OS are comfortable with extrapolated results (± 5%) and associated predicted costs (± 10%) – it must be highlighted that there is potential for error based upon the geographies tested. Throughout the Pilot it was identified that multiple-to-one addresses were most difficult to capture accurately, and whilst the Pilot area contains a set of these (e.g. Leeds City Centre), there are other areas of GB where there is a higher prevalence of these (e.g. Birmingham, Manchester, Edinburgh, London). Any required improvement initiatives to increase OAR accuracy is out of scope of these costs.

7.2. Unique identifier

The Unique Property Reference Number (UPRN) is the unique identifier for every addressable location in Great Britain. It is created by local authorities who have the statutory authority to name and number every street and property and OS who identify objects on the landscape which may otherwise not attract an address. It provides a comprehensive, complete and consistent identifier throughout a property's life cycle – from planning permission or street naming through to demolition.

The UPRN acts as a golden thread, linking multiple information sets about each spatial address in Great Britain. In the same way that each person has a National Insurance number or every book features an ISBN reference, a UPRN uniquely and definitively identifies every addressable location in the country. The UPRN is already used by organisations to link multiple datasets together and to reduce errors in data exchange between each other. For example, a local authority and utility company can continue to hold their own address information in existing formats, but by adding a single field for the UPRN, then can easily link matching records in their disparate databases together.

This Pilot has shown that it is possible to allocate a UPRN to every address within the OAR. The accuracy of UPRNs assigned to one-to-one addresses, OWPAs and LA unique addresses (24.9m records) is 100%. However, for multiple-to-one addresses, 7.6m records, the mathematical probability of that UPRN being accurate is 22.3%. This could result in 5.9m records being allocated the wrong UPRN – thus making the UPRN no longer viable as the Unique Identifier for addresses in GB.

7.3. Potential improvements

There are steps which could be taken to improve the accuracy and completeness of the dataset, such as increased investment in the capture of other OS datasets, i.e. OS MasterMap Topography Layer and ITN Layer.

However, it should be highlighted that this Pilot has demonstrated that whilst the results in creating an OAR are encouraging, one critical issue is the linking of the created addresses with the existing unique identifier, the UPRN. The UPRN has been adopted across Local and Central Government as the key for linking address datasets. The only methodology available for appending a UPRN to an OAR address, results in a potential UPRN error in 18% (5.9m) of OAR records.

It should also be noted that whilst the results of the prototype, when extrapolated up to a national scale, were encouraging, at 96% completeness and 90.8% accuracy, there are various things to consider:

1. 96% completeness does not provide a Universal service – i.e. 1.3 million addresses would be missing across GB;
2. 90.8% accuracy of the address register would result in approximately 2.9 million complete addresses being incorrect. These 2.9m could not be identified prior to the product release, without significant quality assurance investment, such as checking large volumes of records manually, Without this, confidence in the product would be seriously undermined by 1 in 10 addresses being incorrect, affecting usability, especially by organisations such as emergency services and other customers where accuracy is key.

7.3. Usability of the OAR Pilot data

Based upon the evidence contained within this paper, it is the recommendation of OS that this OAR Pilot is not taken forward as a potential solution [redacted] This recommendation is based upon:

1) The lack of a postcode, required for uptake by citizens and users (both public & private organisations)
2) The completeness levels (96%) resulting in a non-universal service
3) The accuracy levels of the completed records (90.8%) resulting in too many errors (2.4 million) to be accepted as definitive / usable by customers for their business applications
4) The existing UPRN would no longer be viable
5) Significant confusion in the marketplace due to the creation of a second spatial address dataset – something that was overcome, after many years of consternation, by the creation of AddressBase.

A ANNEX - SPECIFICATION

COLUMN NAME	COLUMN TYPE	MULTIPLICITY	EXAMPLE	NOTES
OAID	NUMBER	1	1	Unique key
UPRN	NUMBER (12)	0..1	10001071493	Optional as for blocks of flats it might not be possible to assign a UPRN
PARENT_UPRN	NUMBER (12)	0..1	10001071492	To fulfil requirement for parent / child relationships
X_COORDINATE	NUMBER (8,2)	1	518925.74	British National Grid X coordinate
Y_COORDINATE	NUMBER (9,2)	1	204156.69	British National Grid Y coordinate
LATITUDE	NUMBER (9,7)	1	51.723535	ETRS89 Latitude Coordinate
LONGITUDE	NUMBER (8,7)	1	-0.2796186	ETRS89 Longitude Coordinate
ORGANISATION	VARCHAR (100)	0..1	Tesco	Population rates would be expected as very low, and completion only conducted by the Field.
SUB_BUILDING_NAME	VARCHAR (150)	0..1	Flat 2	
BUILDING_NAME	VARCHAR (150)	0..1	Campbell House	
BUILDING_NUMBER	NUMBER	0..1	10	
STREET	VARCHAR (200)	1	High Street	
ALT_STREET	VARCHAR (200)	0..1		Alternative language value of STREET
LOCAL_AREA_NAME	VARCHAR (200)	0..1	Upper Parkstone	Hamlet, village, or possible local name. Milborne St Andrew, Upper Parkstone, Ashley Cross
ALT_LOCAL_AREA_NAME	VARCHAR (200)	0..1		Alternative language value of LOCAL_AREA_NAME
LOCAL_AREA_NAME_2	VARCHAR (200)	0..1		Only to be used if all Local Naming cannot be inserted into LOCAL_AREA_NAME

ALT_LOCAL_AREA_NAME_2	VARCHAR (200)	0..1		Alternative language value for LOCAL_AREA_NAME_2
ADMIN_AREA	VARCHAR (30)	1	Southampton	The responsible authority for the street the address record resides upon.
WARD	VARCHAR	0..1	Parkstone	Might be a nice to have, but depends on quality of NAMED EXTENTS
PARISH	VARCHAR	0..1	St Aldhelms	Might be a nice to have, but depends on quality of NAMED EXTENTS
COUNTY	VARCHAR (50)	0..1	Dorset	Might duplicate ADMIN_AREA, but will give the Ceremonial Country boundary the address falls within.
COUNTRY	VARCHAR (25)	1	England	Country the address resides within,

For any Street or Local Area Names in Wales the Welsh name will be used in the main column e.g. (STREET), the same will also apply for Scotland.

B ANNEX - POTENTIAL THIRD PARTY DATASETS

Please note that 3rd party datasets have not been used at any stage of the production process. Only OS captured data has been used within the method.

Dataset	Source	Link/Website	Contents
ONS NSAL	Office of National Statistics	http://www.ons.gov.uk/methodology/geography/geographicalproducts/nationalstatisticsaddressproducts	UPRN to multiple boundary lookup
Station Usage Estimates	Office of Rail Regulations	http://orr.gov.uk/statistics/published-stats/station-usage-estimates	Location (x,y of railway stations in UK)
Geolytix Code-Point Polygons	GeoLytix	http://geolytix.co.uk/?s=code+point	2012 Open Code-Point with Polygons dataset
ONS NSPL	Office of National Statistics	https://data.gov.uk/dataset/ons-postcode-directory-uk-feb-2016	Lookup from postcode to best fit ONS boundaries
Open Addresses Dataset	ODI, Open Addresses	https://alpha.openaddressesuk.org/developers/apis-and-data	List of addresses captured by ODI
Food Standards Agency	Scores on the doors – food hygiene ratings	http://ratings.food.gov.uk/open-data/en-GB	Food hygiene ratings of commercial food outlets and their address
CQC	List of healthcare practitioners	http://www.cqc.org.uk/content/how-get-and-re-use-cqc-information-and-data#directory	List of healthcare practitioners and their address
OFSTED	Schools addresses and names	https://www.compare-school-performance.service.gov.uk/download-data	Schools addresses and names
VOA data	Classifications		
GP Practice data	HSCIC	https://data.gov.uk/dataset/england-nhs-connecting-for-health-organisation-data-service-data-files-of-general-medical-practices	List of all GP practices in England and Wales. Scotland?

Dataset	Source	Link/Website	Contents
Libraries Dataset	Collections Trust	https://data.gov.uk/dataset/uk-public-library-contacts-14032012	List of all UK libraries
Listed Buildings	Historic England	https://historicengland.org.uk/listing/the-list/data-downloads/	Parks, Gardens, Listed buildings, scheduled monuments and other datasets.
National Public Transport Gazetteer	Department for Transport	https://data.gov.uk/dataset/nptg	PTG is a database of localities (cities, towns, villages and other settlements) in Great Britain.

C ANNEX – MULTIPLE-TO-ONE ADDRESS UPRN ASSIGNMENT

THE PROBLEM

In buildings with multiple-to-one address UPRNs have been assigned to the addresses randomly. It is required to work out the impact of this random assignment, by deriving a quality value of the percentage of UPRNs expected to be correctly assigned.

DISTRIBUTION OF UPRNS

- **Single occupancy**

We assume that every single-occupancy building is assigned the correct UPRN. So the probability of a UPRN in a single-occupancy building being correct is 1.0, or 100%

- **Double occupancy**

In a double-occupancy building, the true UPRNs are A and B, in that order, denoted by [AB]. The possible permutations of randomly assigned UPRNs are {[ab] and [ba]}. Assuming the 2 permutations are equally likely (i.e. we could equally well have assigned [ab] as [ba]) then the following probabilities can be calculated:

P(0 correct) = {[ba]} = 1 out of the two possible = 1/2 = 0.5

P(1 correct) = {this never happens} = 0 out of 2 = 0

P(2 correct) = {[ab]} = 1 out of the 2 possible = 1/2 = 0.5

To calculate the mean number of correct UPRNs, multiply each of the numbers [0,1,2] by its respective probability [0.5,0,0.5] and add them all up:

The mean number correct = 0*0.5 + 1*0 + 2*0.5 = 1.

This value (1) makes sense, as half the time you will have 2 correct and the other half of the time you will have 0 correct, so on average you will have 1 correct.

- **Triple occupancy**

We now have three UPRNs, denoted [ABC]. The possible permutations of randomly allocated UPRNs are the six outcomes: {[abc], [acb], [bac], [bca], [cab], [cba]}. The probabilities are:

P(0) = {[bca],[cab]} = 2 out of 6 = 0.33333

P(1) = {[acb],[bac],[cba]} = 3 out of 6 = 0.5

P(2) = {} = 0 out of 6 = 0.0

P(3) = {[abc]} = 1 out of 6 = 0.16666

The mean number correct = 0*0.33333 + 1*0.5 + 2*0.0 + 3*0.1666 = 1

In other words, if you have lots of buildings with 3 UPRNs, on average you will get one UPRN correct per building.

- **Quadruple occupancy**

Using the same notation, we have the true UPRNs denoted by [ABCD]. The 24 permutations are:
{[abcd],[abdc],[acbd],[acdb],[adbc],[adcb],

[bacd],[badc],[bcad],[bcda],[bdac],[bdca],

[cabd],[cadb],[cbad],[cbda],[cdab],[cdba],

[dabc],[dacb],[dbac],[dbca],[dcab],[dcba]}

The probabilities are:

P(0) = {[badc],[bcda],[bdac],[cadb],[cdab],[cdba],[dabc],[dcab],[dcba]} = 9/24 = 0.375

P(1) = {[**a**cdb],[**a**dbc],[bca**d**],[bd**c**a],[cab**d**],[c**b**da],[da**c**b],[d**b**ac]} = 8/24 = 0.333333

P(2) = {[**ab**dc],[**a**cb**d**],[**a**d**c**b],[ba**cd**],[c**ba**d],[d**bc**a]} = 6/24 = 0.25

P(3) = {} = 0.0

P(4) = {[**abcd**]} = 1/24 = 0.0416666

The mean number correct is 0*0.375 + 1*0.333333 + 2*0.25 + 3*0.0 + 4*0.416666 = 1

The outcome is one again. In fact, if you do the same exercise for any number of permutations, the mean number of correct UPRNs always works out as exactly 1.0.

CALCULATING THE PERCENTAGE OF CORRECTLY ASSIGNED UPRNS

Now we have ascertained that the mean number of correctly allocated UPRNs is always 1, no matter how many UPRNs are in the building, the calculations become simple. The mean number of UPRNs that are correct equals 1.0 times the number of buildings, which of course is just the number of buildings. Say we had the following example:

No. UPRNs per building	No. of occurrences	No. of UPRNs	Estimated No. of UPRNs correct
1	100	100	100
2	20	40	20
3	10	30	10
Total	**130**	**170**	**130**

In the above example, the percentage of correct UPRNs is:

100 * (total no. correct)/(total no. of UPRNs)

100 * 130/170 = 76.4%

CALCULATIONS SUING THE REAL DATA

If we perform the same calculations on the actual numbers of buildings with single and multiple-to-one addresses, we get:

Percentage of correct UPRNs in the real dataset = 75.9%

Ignoring the single occupancy buildings, which are always correctly assigned, we get:

Percentage of correct UPRNs in multiple-to-one buildings in the real dataset = 22.3%

CAVEAT

The caveat to the above is that, for buildings with large address-counts, there are very few occurrences (only one in many cases), while the statistics assume that there are many occurrences. This has not been taken into account.