

NON ROYAL MAIL IP OPEN ADDRESS REGISTER: PILOT INTERIM REPORT

RESPONSIBILITY FOR THIS DOCUMENT

████████████████████ (Ordnance Survey) are responsible for the content of this document.

PUBLICATION DATE

Draft Version - July 2016

CONTENTS

NON ROYAL MAIL IP OPEN ADDRESS REGISTER: PILOT INTERIM REPORT	1
RESPONSIBILITY FOR THIS DOCUMENT	1
PUBLICATION DATE	1
CONTENTS	1
1. EXECUTIVE SUMMARY	2
2. PILOT PROJECT	3
A. BACKGROUND	3
B. AIMS	3
C. ASSUMPTIONS	3
D. CREATION METHODOLOGY OUTLINE	4
E. QUALITY ANALYSIS METHODOLOGY OUTLINE	4
3. PILOT MANAGEMENT	5
A. FORM OF PILOT	5
B. TIMELINE	5
4. CREATING CONTENT: DETAIL	6
A. STAGE 1: GENERATE CANDIDATE LIST	6
B. STAGE 2 - AUTOMATED GEO-PROCESSING	7
C. STAGE 3 - MANUAL GEO-PROCESSING	10
D. STAGE 4 - FIELD GEO-PROCESSING	11
6. CONCLUSIONS TO DATE	12
7. NEXT STEPS	13
A ANNEX A – SPECIFICATION	14
B ANNEX B – POTENTIAL THIRD PARTY DATASETS	16
C ANNEX C – AUTOMATION ACCURACY ERROR EXAMPLES	18

I. EXECUTIVE SUMMARY

Ordnance Survey (OS), on behalf of The Department for Business, Innovation & Skills (BIS), is exploring options for the creation and maintenance of an address register which does not contain any intellectual property rights of Royal Mail Group Limited (RM).

This report acts as an interim update on the progress made by OS on the Pilot Project. This is based upon OS working for 7 weeks, since 23rd May 2016, on prototyping the methodology outlined in the Non RM-IP Address Register Solution Paper.

The aim of the Pilot Project is to deliver improved confidence and evidence as to the expected quality of the address records (accuracy and completeness) at each stage of the creation process.

Using a variety of methods and content, as explained later in the document, OS can report that at this interim stage, the following has been achieved:

1. Creation of product specification (as per Annex A)
2. Capability developed to carry out the pilot project
3. Generation of approximately 35million candidates which require addresses, i.e. UPRN, USRN, X, Y
4. Creation of 8.1m complete addresses with an accuracy of 91.3%, e.g. "4, Rhodfa Fuddug, Aberystwyth"
5. Creation of 16.8m incomplete addresses where 2 or more Primary fields are populated, e.g. "14, High Street, Hampshire"
6. Creation of 9.9m incomplete addresses where 1 Primary field is populated, e.g. "12" or "High Street"
7. Identification of a further 2.7m full addresses which can be directly ingested from NAG which are currently contained within either 16.8m / 9.9m incomplete addresses, e.g. "38 Kinmel Street, Liverpool"
8. Identification of 3rd party datasets that it is believed do not contain Royal Mail IP that would be of value to the Pilot Project.
9. Manual assessment of 54,000 addresses.
10. Field based assessment of 3483 addresses.

Whilst the completeness and accuracy achieved to date is promising, we believe there is value in continuing the pilot for the following main reasons:

1. Transition as many as possible of the remaining incomplete addresses into complete addresses through automation, manual and field geo-processing
2. Better understanding of the completeness and accuracy of desk based and field based geo-processing on different address types
3. Investigate the ingestion of OWPA's (estimated to be 3 million) into the OAR
4. Report on the feasibility and quality of a national RM IP free OAR
5. Commence formal engagement with ODI based on agreed remit between GDS and BIS. OS / ODI engagement not been able to commence to date due to HMG and ODI unable to agree work arrangements.

2. PILOT PROJECT

A. BACKGROUND

Further to the statement in the 2016 Budget regarding an open address register, BIS has 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 have proposed in outline, how such a register could be created and maintained within the solution paper “Non RM IP Address Register”. Subsequently OS has been requested 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 has agreed to undertake such a pilot project.

B. AIMS

The pilot is required to prototype 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 will be demonstrated.

The pilot will focus on a statistically significant sample of properties covering a range of address types in a variety of geographies. This is to ensure that any issues associated with different address types and geographies are encountered.

The methodology that will be tested is outlined under the Pilot Project Methodology section.

The parties agreed that OS will deliver an interim report summarising the results of the initial phase of the pilot, comprising the methodology and completeness of Stage 1 and 2 methodology. This document is this deliverable.

C. ASSUMPTIONS

The parties agreed that the pilot will 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 will not be used in the initial capture, although such data will be investigated for its potential to be used in the creation and maintenance of the Solution.

[REDACTED]

The complete Solution will cover addressable properties¹ within Great Britain. To be clear, the Solution will not contain a postcode. As a minimum, every record in the Solution should contain the following:

- 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

¹ An addressable property means anything that attracts a rateable value. It would include electricity sub stations and billboards but not geographical features such as ponds or cattle grids.

D. CREATION METHODOLOGY OUTLINE

The high level methodology being tested by the Pilot Project is as follows. Each stage is described in further detail within section 4.

STAGE 1 Generate Candidate List	➤ Creation of an address candidate list from the NAG which will contain the UPRN, USRN and X/Y coordinate
STAGE 2 Automated geo-processing	<ul style="list-style-type: none"> ➤ Sophisticated geo-processing to identify the building name/number from OS data and other open sources ➤ Use of OS products to assign an address locality (street, town, ward, county)
STAGE 3 Manual geo-processing	<ul style="list-style-type: none"> ➤ Manual intervention in scenarios where automated processing cannot identify a building name / number or locality ➤ Use of OS data and investigation of 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

E. QUALITY ANALYSIS METHODOLOGY OUTLINE

Reporting on the quality of addresses at each production stage is essential for the success of this pilot. Both the completeness and accuracy of the OAR will be 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 need to be populated in order to create a “complete” address. We have identified the essential primary fields required to make a complete address and calculated the total volume of records for partial and complete addresses.

➤ **Comparison to AddressBase**

To determine accuracy, the OAR records have been compared with AddressBase in order to examine where they are different. The AddressBase product has been used as a baseline dataset.

➤ **Differences between OAR and AddressBase**

In circumstances where the OAR is found to be different to AddressBase, manual analysis has been conducted to assess whether the address is still correct or not.

➤ **Address field analysis**

Address quality is measured at a number of levels including the individual primary fields, by comparing these to AddressBase. This identifies common trends and enables the teams to act on the findings.

➤ **Stratified data samples**

Different geographical environments present different challenges when populating addresses. We would expect to see a variation in the success rate, assessment time and quality between different areas. Therefore it is vital to ensure that a good stratification of these environments are assessed so that conclusions can be accurately made.

3. PILOT MANAGEMENT

A. FORM OF PILOT

The pilot has been established as a project and initial set up administration has been completed with the stages of the methodology being assigned as work streams. Each of the work streams have defined and developed methodologies and begun work on populating the address register.

B. TIMELINE

The timeline in Figure 1 displays the schedule of work packages.

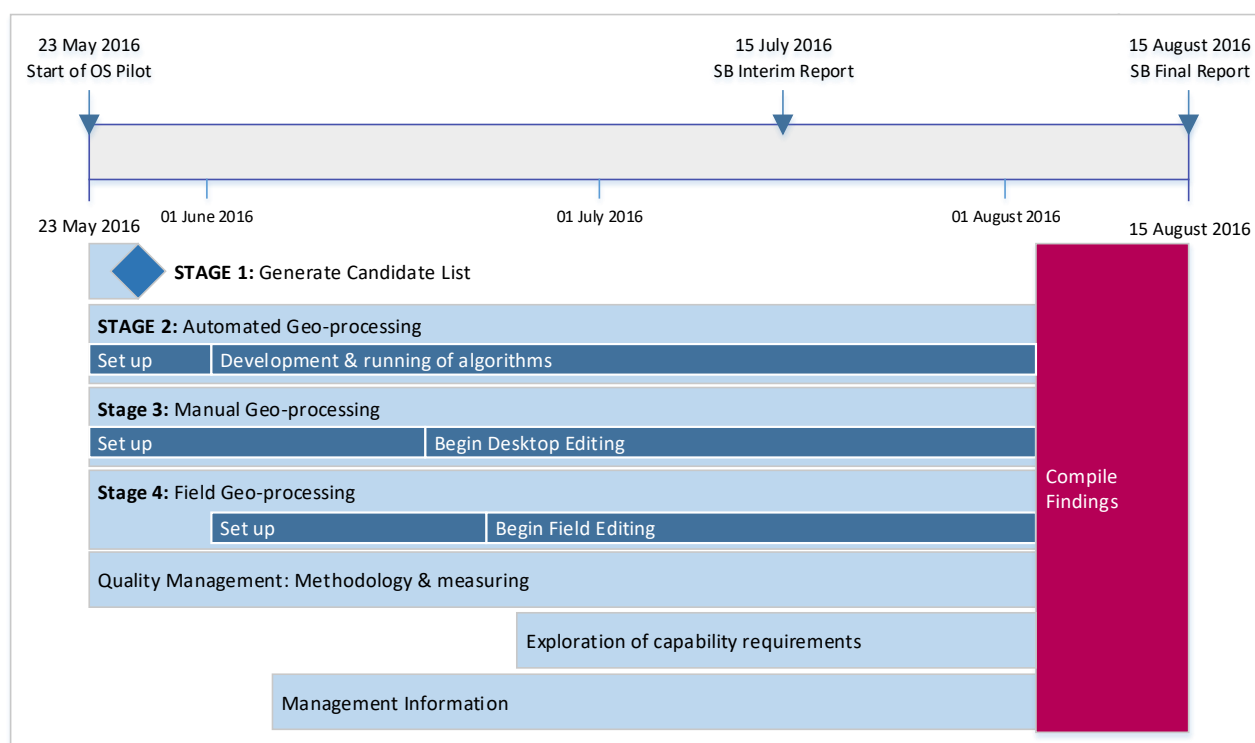


Figure 1: Project Timeline

4. CREATING CONTENT: DETAIL

A. STAGE I: GENERATE CANDIDATE LIST

I. METHOD

A product specification has been developed to enable a 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 records² achieved using a flag inserted by Local Authorities.
2. Filtering of the above records undertaken by using a classification code inserted by local authorities. This meant items such as Street Records and other addresses out of scope could be removed.
3. Objects Without a Postal Address³ (OWPAs) were then identified using a defined list of classifications.

II. RESULTS

Approximately **35 million** candidates were selected using the process above.

The address specification and definitions have been further defined and developed, in particular to now include OS OWPAs (e.g. ponds & cattle grids), hence the increase from the OS original estimate of 32 million records.

III. 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.

Within the manual and field geo-processing stages of production there may be circumstances in which addresses are identified that are 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 manual and field teams would capture the additional units and therefore the total number of records would be increased.

IV. 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.

² Approved / Live records – These are all addresses which a Local Authority has marked as currently existing.

³ Objects Without a Postal Address – 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 things such as Ponds and Electricity Sub Stations.

B. STAGE 2 - AUTOMATED GEO-PROCESSING

I. METHOD

Using the Specification (Annex A), the Automation Team have focussed on populating primary fields of an address for each candidate, in particular the Building Name, Building Number and Street Name, which were expected to be the most challenging fields to automatically identify. The algorithms developed have been run nationally and due to the complexity in some cases have taken up to 1 week to return results. The following data sources have been analysed and interrogated to extract the required data:

1. Use of OS core content

Using OS data sources, the automation team were able to assign building names and numbers to Building TOIDs⁴. This has currently only been completed for one to one relationships. This means where OS has a building name or number only attached to one building with one address seed contained within it.

2. Use of OS Mastermap Highways production system

OS has recently released a beta product to the Public Sector community which is an amalgamation of OS Mastermap Integrated Transport Networks (ITN) product and the NSG (National Street Gazetteer). This means using the Unique Street Reference Number (USRN) which is contained within the candidate list; a street name can be returned for the address records where OS has been able to match its ITN product spatially to the NSG. Please note this method can only be used for England and Wales due to coverage and has currently only been implemented where OS has a singular name for a given USRN.

3. Use of OS Mastermap Integrated Transport Networks (ITN) and spatial interpolation techniques

Where a street name could not be determined using the above method, ITN and the spatial function of 'Find My Nearest' was used to try and allocate a street name. This has currently only been explored for Scotland as the above technique has yet to be exhausted for England and Wales. This technique will also be refined over the course of this pilot as it is yet to be fully exploited or tailored meaning only small numbers of addresses have had their street allocated under this method.

4. 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 a number of different ways and advanced GI techniques; firstly grouping together candidates of addresses, and then using the already assigned building numbers to extrapolate additional building numbers assigning these to address candidates via the Building TOID.

5. Local Authority only addresses with no RM equivalent

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.

II. RESULTS

As described under the Method, a number of techniques have been run in parallel in order to populate different elements of an address. There is a degree of crossover between methods making it difficult to report on output volume per methodology. Therefore we have reported on the total output of the automation stage as a whole.

⁴ 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.

➤ COMPLETENESS

Figure 2 demonstrates the progress made on populating addresses for each of the candidate records. We can see that 8.1 million records have been resolved and have all required primary fields completed. A further 16.8million records have 2 primary fields completed and 9.9 million records have 1 primary field populated. As the focus of the automation team has been on populating the Building Name, Building Number and Street Name it is not surprising that a large portion of the records have some but not all attribution completed. It is expected through the population of the village / town/ city name we will see a significant increase in the population of all primary fields. It is likely that the village / town / city will be a relatively straight forward query to complete. The greater challenge will be with sub building naming and numbering and multiple occupancy properties. There are also geographical differences between datasets that we are using which may cause challenges going forward. For example as described in method 2 OSMM Highways dataset, data is only available in England and Wales.

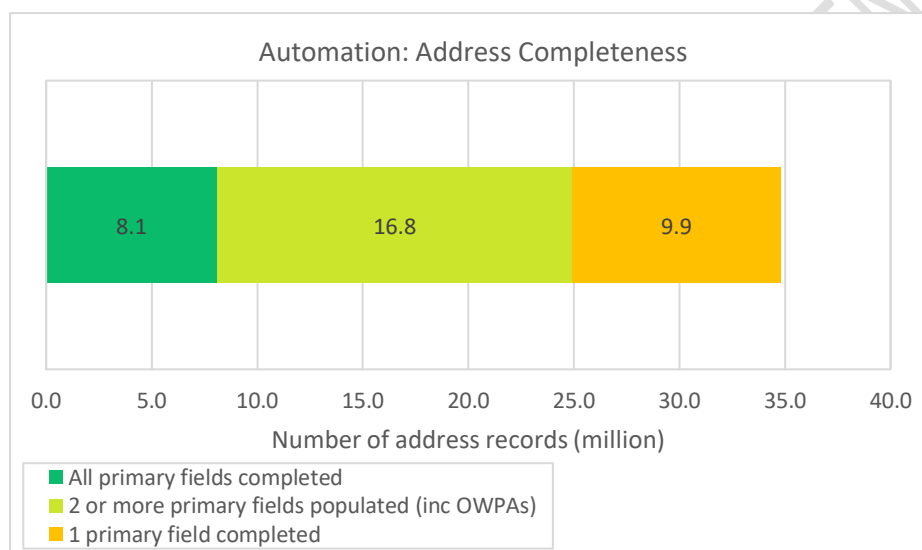


Figure 2: Breakdown of results from the automation stage.

➤ ACCURACY

To determine the accuracy of the completed automation results, records with all primary fields completed (8.1 million) were compared with AddressBase. This was achieved by counting the number of characters that were different between the two records. Where they were different, a manual investigation was conducted as to why and what errors, if any, were present.

Figure 3 displays the comparison between the automated results and AddressBase. It was found that 85% of records matched AddressBase. Of the remaining 15%, just under half were found to be valid addresses and the remaining were incorrect. The most common example of addresses being different yet valid in both instances are where there are minor spelling and grammar differences in the street names.

Figure 3: Automation accuracy results

Address match ⁵ between automation output & AddressBase	85.0%	6.9 million records
Address different to AddressBase but still valid	6.3%	510,300 records
Address different to AddressBase and found to be incorrect	8.7%	704,700 records

⁵ A match is defined as a difference of less than 6 characters between the two addresses. Analysis has shown that in most scenarios less than 6 characters accounts for abbreviations and differences in punctuation, but the address is essentially the same.

The cause of errors has been analysed and there are number of reasons why the automated process may have produced an incorrect address. The most common cases are for Closes and Courts (where the address has been assigned to the nearest road and not the court the address sits within) and instances where there were multiple terraces with the same numbering sequence on the same road. Figures 4 provides graphical representation of the issue and a description. For further examples of quality errors identified please see Annex C.



Figure 4: Example of an incorrect street name applied to an address

As part of the automated process, the address (the cyan dot) has been assigned 'Elwick Avenue' as its road. However it is actually a member Claxton Court. The name Claxton Court isn't recorded in the servicing road and so it will never be assigned automatically. This is a problem as there is already a legitimate '12 Elwick Avenue' further down the road. Courts and Closes have proved a consistent problem for the automated matching.

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. This would not be such a problem if it were not for the fact that these records have been flagged as complete by the automated process (there is nothing in the data to easily pre-empt these sites). A solution to identify these sites is being investigated.

The quality completeness and accuracy analysis of automated address records has been an essential feedback loop for the automation team. We have a clear understanding of the output and where further work is required, and can be confident in the completed records to date. The current results do not indicate the full capability of automation, however demonstrate the levels of accuracy that could be achieved. Of the completed output our findings show that 91.3% of populated addresses match AddressBase and or are correct. This will not be true for all types of Address though, and we expect the levels of accuracy to vary depending on the address geography, data source, automation method and address type.

III. TIME

A total of 7 weeks has been focussed on the automation stage. Although good progress has been made, there are further quick wins that could improve the volume and quality of the output. The automation team will continue to develop their methods until the end of the trial, at which we will have a view of what could be done further in a real production scenario.

C. STAGE 3 - MANUAL GEO-PROCESSING

Manual intervention is required in scenarios where automated processing cannot identify addresses or certain aspects of addresses. The most efficient 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.

I. METHOD

The infrastructure has been established to enable content to flow directly from the automation databases to the teams conducting the manual geo-processing. So that the team can complete a statistically significant sample of data within the timeframe of the pilot, work has begun on manual editing in parallel with the automation development. The automation element of the pilot have found this to be beneficial as lessons can be fed back and processes improved, particularly in circumstances where an address as been incorrectly populated.

To aid with the decision making process of populating an address, the manual edit team currently have access to the following datasets:

- OS Imagery
- OS Maps API service
- Road Link Layer (Containing USRN attribution sourced from Highways)

II. RESULTS

When identifying priorities of where to focus the manual production effort, it was important to get a good stratification of urban and rural, residential and industrial environments. These each present different challenges to address capture and subsequently we would expect 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>

Where possible, the priority areas were selected to provide a representative sample for addresses across the country. Beyond that they were kept as contiguous as possible in order to aid production efficiency and help produce a viable product sample.

To date 54,000 records have been assessed. The assessment time per address will vary greatly depending on the type address, some will be more challenging than others.

On an initial sample of 5,385 populated addresses by the manual team, 94.2% were found to match AddressBase. Approximately half of the results that did not match AddressBase were found to be valid. As described in section B this is commonly due to spelling and grammar differences in the street names.

For efficiency reasons, we would not recommend that the processes run in parallel in a real world production environment.

III. TIME

The time required to complete the manual geo-processing effort in a real production scenario will depend on the success of the automation output. Our next step within the pilot is to ensure we have an appropriate stratification of property types for the manual team to assess. This will enable results and rates to be extrapolated nationally.

D. STAGE 4 - 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. Although this a highly effective method, it is time consuming and does not always result in a resolved record.

I. METHOD

Our field resources are able to access data provided by the manual geo-processing team. They visit a property on the ground and populate missing fields. Within urban areas where candidates are within close proximity this can be a quick process. However challenges present themselves in rural and industrial areas. Photographic evidence has been gathered by the field team to aid with further quality discussions and to support the other stages of production.

II. RESULTS

The field team are in the early stages of production and have assessed 3483 address records. They have been able to populate 89% of the 3483 candidate addresses. The accuracy of these populated addresses has not yet been analysed, however due to the methodology it is expected to be high. The areas of assessment have mainly been residential and commercial.

Anecdotal evidence has shown that the majority of unresolved addresses has been due to not being able to determine the building name or number. As more properties are assessed we will be able to report and provide examples of where an address cannot be resolved.

It is worth noting that field team have captured organisation names with a high degree of accuracy and are the only production process to have done so. The field team were also able to correctly identify street names (in situations where the manual team have made an error on the same street) and were able to better identify property group names, such as named terraces or courts.

III. TIME

It is too early within the pilot to determine a rate of effort due to the low number of records assessed and limited variety of property types.

6. CONCLUSIONS TO DATE

The generation of a candidate list stage 1 was completed quickly and a specification has been defined. Stages 2, 3 & 4 have completed their set up and begun to process addresses.

In order to refine the methodology and quality of results, the automation stage has continued to be developed throughout the pilot. The manual and field geo-processing work streams have run in parallel and provide valuable lessons and feedback to the automation stage.

To date out of a potential candidate list of 35million records, 8.1 million have been fully populated, 16.8 million have 2 or more essential fields populated and 9.9 million records have 1 essential field populated. These results are wholly from the automation stage of the method and with further developments in the pipeline we are confident in increasing the success rate significantly.

We have been able to focus on a national coverage for the automation results, however for the manual and field efforts we will use statistically significant stratified sample of records and extrapolate where appropriate. Production stage 3, manual geo-processing, have assessed 54,000 candidates and stage 4, field geo-processing, have assessed 3483 records. It should be noted that both stages 3 and 4 have only begun populating candidates for a short period of time and productivity is planned to increase in the final half of the pilot.

The results to this point within the pilot have been positive and the teams have made good progress. Much of the first half of the pilot has been focussed on defining and setting up the production methodologies and understanding what is possible. We have a clear plan for the final stages of the pilot and will target areas where we expect the population of address to be challenging, for example industrial estates and shopping centres. This will be invaluable to our understanding of how complete and accurate the RM IP free OAR could be, and will validate the costs and timescale associated with creating an OAR. We are currently not in a position to provide an update on the costs associated with creating a RM IP free OAR. The final volume of addresses populated by the automation stage and work rates from the manual and field geo-processing will be used in the end stages of the project to re-assess costs.

7. NEXT STEPS

Whilst we have had some success to date, there is still far more that can be improved and tested. As a summary of intent, OS's next steps would be:

- **Develop and expand the stage 2 automation algorithms**

Further development is planned for the methods described within the stage 2 methodology as well as new ideas. It is anticipated that these developments and new ideas will increase the volume and quality of populated addresses. Examples include:

- Expansion of the use of OS content to include where multiple to one relationships exist – for example between an OS street name and a USRN in the NAG hub and also where OS content has more than one name or number for a given Building TOID.
- Extensive expansion in using GI techniques to increase the number of street names assigned
- Expansion of the use of AddressBase Lite records, specifically in relation to Multi-Occupation buildings
- Deriving address elements for OWPA records
- Investigating the automatic interpolation of sub building names / numbers for Multi Occupancy buildings
- Work on higher level geographies e.g. Locality and Town Names
- Feedback from the Manual and Field geo-processing teams on incorrectly populated addresses
- Increase the number of Primary fields populated to move more addresses to 'Complete' status. We are confident that a minimum of 3 million address would be resolved instantly by assigning a town name and there are likely to be many more

- **Expand the manual and field geo-processing effort**

The stage 3 manual geo-processing and stage 4 field geo-processing teams are confident in the methodology that they have implemented and that staff have achieved the required quality assurance during training. The focus for the teams for the remainder of the trial will be to ensure that an appropriate sample of all address types have been assessed. This is vital to enable results and rates to be extrapolated nationally and quality completeness and accuracy to be correctly measured.

- **Investigate "building capability" costs and timescales**

Further investigation is required to understand the infrastructure, technology, tools and skills needed to implement each stage of the solution methodology. The process by which data is ingested by GeoPlace would also require development and the scale of this needs to be investigated. The publication capability must be considered, in particular, whether the current publication platform could be adapted rather than a whole new capability built.

- **Develop an understanding of the impact on the maintenance processes**

Third party data has not been included in any of the automation techniques, however a number of potential sources have been identified (Annex B). These may improve data quality and completeness levels, but further analysis is required to assess the value and legitimacy of use in the OAR (i.e. that they are RM IP free and open).

A ANNEX A – 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 GDS (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 B – 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/metadata/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	GeoLytx	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?

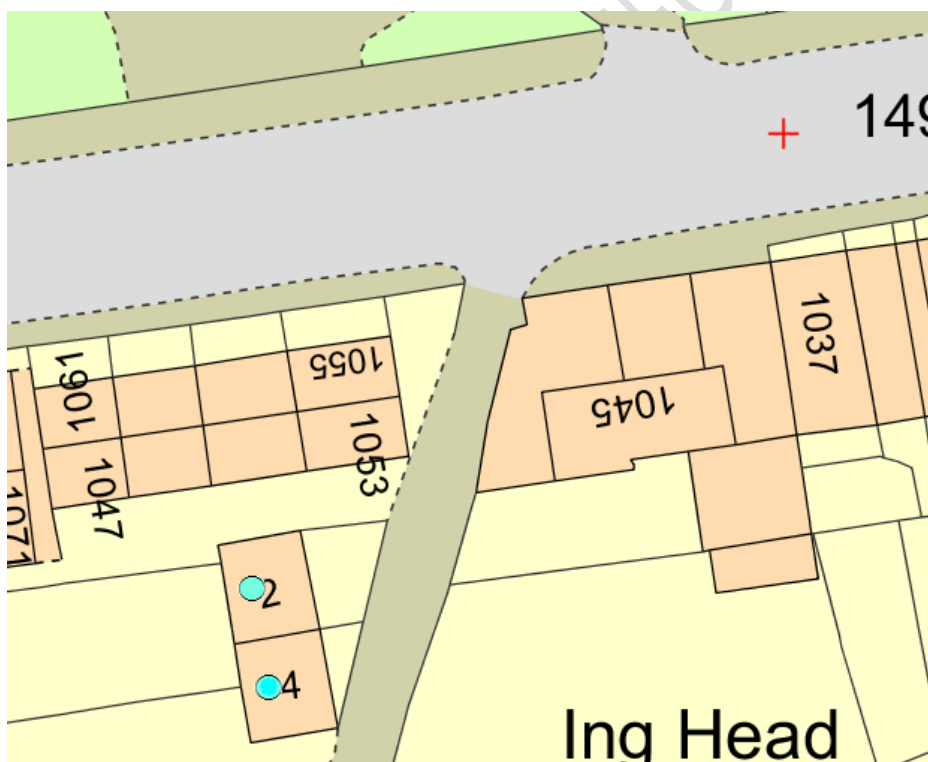
Dataset	Source	Link/Website	Contents
Libraries Dataset	Collections Trust	https://data.gov.uk/dataset/uk-public-library-contacts-14032012	List of all UK libraries
Listed Buildings	Historic England	https://historicengland.org.uk/listing/the-list/data-downloads/	Parks, Gardens, Listed buildings, scheduled 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 C – AUTOMATION ACCURACY ERROR EXAMPLES

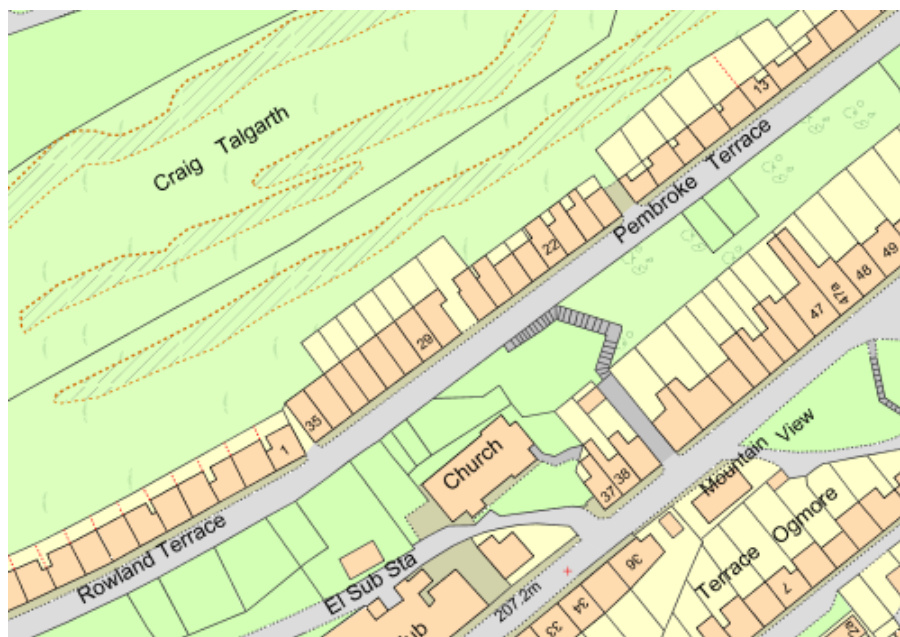
Example 1: In this example the differences are very minor. The Open Address Register record is captured as 'Penrhyn, Bron-Y-Felin road' whereas AddressBase has 'Penrhyn, Bron Y Felin road' captured. Both may be considered correct. These sort of minor differences (changes in grammar, spelling mistakes in AddressBase and the lengthening of abbreviations) account for majority large number of differences in the existing datasets.



Example 2: These addresses 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.



Example 3: 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 below) 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.



Example 4: One of the most common examples of the Open Address Register address being significantly different to the AddressBase addresses, yet still both addresses being correct is where the OAR has captured the road and AddressBase hasn't. The example below is on a military base where the AddressBase records were simply the house number, followed by Oxendene and the post code. Another common example is where AddressBase describes the road geographically but OAR is able to give the proper name i.e OAR: Old School Terrace versus AddressBase: U3357 Spittal Green to JCT B4329 Froghall Cross.

