MAYOR OF LONDON

SAFESTATS

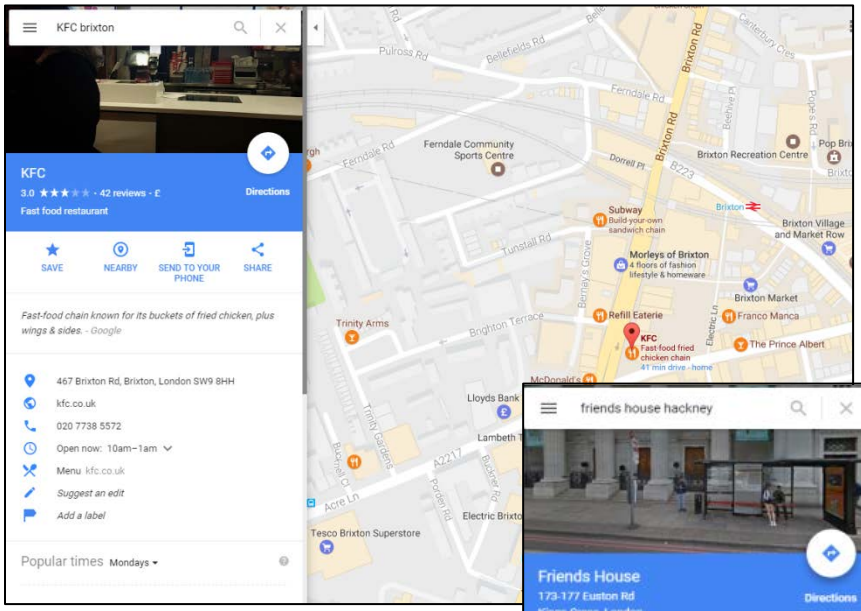
INFORMATION SHARING FOR TACKLING VIOLENCE

Geocoding ISTV data – methodology

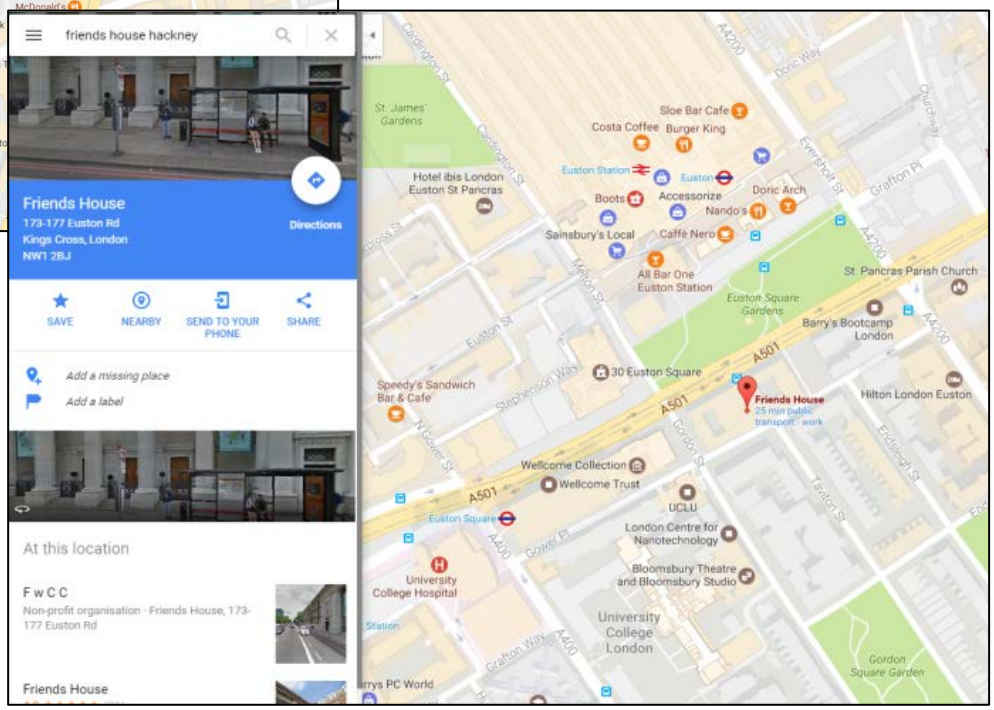
Background

This data is part of the Home Office Information Sharing for Tackling Violence (ISTV) project, where hospitals are encouraged to record additional information at their A&E receptions around the injuries suffered by victims of violence. This is with the aim of then sharing with other public safety bodies to enrich ongoing preventative work and identify new priorities. Of the additional data now recorded at over 25 hospital receptions in London, the location of the violent incident is key to assisting this process.

Due to the understandable pressures existing on victims and reception staff, the freetext locations recorded are often of varying quality, and therefore cannot be mapped using one single process. Whilst it would be helpful to pass all of the location data through a free online geocoding source (for example Google maps or Streetmap), testing has showed how the quality of the data will influence the accuracy of the results, as shown in the images below.



Accurate geocode using Google



Inaccurate geocode using Google

For this reason it is clear that although online resources clearly have use in certain situations, only records with certain information present can be passed to them; with other methodologies required for data that isn't suitable.

Aims

The aims of the geocoding process are four-fold and the methodology is organised in a similar fashion:

1) To standardise and join together data from all providing hospitals into a single dataset;
2) To identify any hospital records that have any incident location information which contains geographic data (eg. a street, a postcode, a town);
3) To assign these records to a suitable level of geography depending on the quality of the information;
4) To make this data (along with other information held in the records such as date, time, providing hospital etc) publically, but securely, available in a map-based interface.

Software used

All processing is carried out in Safe Software's Feature Manipulation Engine (FME) product, and the mapping interface built sing Javascript and ESRI.

Methodology

1. Standardising and Joining the source hospital data

The individual hospital datasets are read in from the SafeStats secure SQL database with their formats required to be standardised in order that they can be appended to one another. There are only 7 key columns that are required for the geocoding outputs:

- Hospital name
- Unique ID of record
- Incident Data
- Incident Time
- Incident Location
- Incident Location Type
- Method of Injury

In most cases these columns and the relevant data already exist for each hospital, however in some circumstances either they do not, or additional columns that may be relevant also exist; therefore a set of decisions have had to be made about how best to populate them.

- If no incident time or date is recorded by a hospital, the arrival time/date is used (required in approximately 4% of cases). If there is still no time or date, a default of

01/04/2009 and 00:00 is used in order that the record can still appear in the webmap output.

- If an incident postcode column is present as well as an incident location, then the two are combined into a single 'Incident Location' field. The same is relevant for extra information about the method of injury (eg. body part injured rather than just the weapon used).

All other columns/fields are then dropped, and the data can then be appended to one another to create a single dataset.

Before the data can be assessed for its suitability for geocoding, the date and time formats require standardising across all hospital data within the single dataset, relevant hospital trusts and injury-type groups need to be added, and extra temporal columns added for future use that split out the dates/times into years, months, days of the week, hours, and hour periods.

2. Assessing the data for geocoding suitability

A large amount of the incident location column contains data not useful for geocoding. In most cases these are merely blank, or contain strings of text such as 'Unknown', 'NK', 'Public Place', 'Nil location' or 'Removed by hospital'.

A thorough search is carried out of the complete source dataset to ensure these records are captured, assigned with 'Insufficient address information provided' and then not taken forward for geocoding. They are however saved for final outputs (see stage 6).

3. Geocoding & assigning relevant geographic levels

i) <u>Cleanse</u>

In order to assign geography to the records, the address elements that are present in the record (eg. postcode, street, town) first need to be identified. This involves removing and/or standardising common terminology and phraseology used by patients and hospital staff in order to simplify the geocoding process, and ensuring that as many records are correctly geocoded as possible.

Examples of this include:

- *police, police stn, pol, custody, nick*, all being amended to show '*police station*'
- *tube, underground, ug, u/ground, u/g, tube stn, u/g stn* (etc) amended to show 'underground station'
- *st and strt* amended to show 'street'
- *traf square, trafalger sq, Trafalgar sqr (etc) amended to show 'Trafalgar Square'*
- *Forest Gt, F Gate* amended to show *'Forest Gate'*
- *'near to', 'behind', 'outside'* being removed as not useful address text

- *@,?, /, £,$,% also removed for the same reasons*

NB. Each of these search terms are surrounded by spaces to ensure characters such as ug and st that exist within words are not picked up.

This cleansing process is based on the records currently held, but will continue to be updated to include any new permutations of phrases seen in future data.

ii) Split

By using a set of regular expressions (regex) to search for a specific set of text strings it is possible to identify which recognisable address elements are contained within the incident location (eg. postcode or street). Where these elements are present, they are split out from the full incident location and placed in a relevant new column (eg. 'full postcode', 'area'). For example:

- does it contain a full postcode?
  Regex: [a-z]+[0-9]+.[0-9]+[a-z]+
- does it contain a partial/district postcode?
  Regex: [a-z]+[0-9]+
- does it contain a known area?
  Compared to a lookup table of over 500 known 'areas' in London
- does it contain a road/street name?
  Compared to a lookup table of 75 recognised road suffixes. Naturally it is not just the suffix that needs to be extracted, but the road/street name too. However this presents a problem as road names can be multiple words eg. high street, upper high street, upper high mill street. In order to cover all of these options, 3 'extract' columns are used instead of one, for the recognised suffix with a single preceding word, two preceding words and three preceding words (where available). These can all then be used in sequence when matching to streets later in the process.

Whilst the above process is a more than suitable way of categorising the vast majority of records, there is a large proportion that contain commonly used locations or location types. These include hospitals, stations (train, tube, bus, police), schools, and London landmarks. By identifying the presence of these, it is possible to specifically geocode these straightaway using in-house lookup tables containing their exact co-ordinates.

A lookup table of over 50 landmarks is therefore used to search for London places of interest, and regular expressions used again in the case of stations, hospitals and schools:

- *Police Station|British Rail Station|Rail Station|Bus Station|Tube|Station*
- *Hospital|Ward|Hosp (etc)*
- *School|College|Academy|Primary|Secondary|Pupil Referral* Unit (etc)

It is accepted that this will never be able to pick up 100% of all these types of locations (eg. St Marys CofE could be a school or a church) however this is a suitable way of ensuring that there is manual and therefore more accurate control over the geocoding of identifiable locations. As noted previously, the lookup tables and regex strings used in this process are based on the records currently held, but will continue to be updated to include any new locations/location types commonly seen in future data.

iii) Prioritise

Having identified the presence of address element(s) found in each record it is then possible to not only group them accordingly for assigning to different geocoding processes, but also to prioritise the combinations of these elements in order of which is most likely to give an accurate point on a map. For example:

A location containing a:

Hospital/Station/School/Landmark
has a better chance of a point on a map than one containing a
Full postcode
than one containing a
Street+Partial Postcode+Area only
than one containing a
Street+Partial Postcode only
than one containing a
Street+Area only
than one containing a
Street only
than one containing a
Partial postcode+Area only
than one containing a
Partial postcode only
than one containing an
Area only
than one containing
Nothing

Having assigned every record one of these combinations, they are ready to be split out and geocoded in specific ways depending on their assignation.

iv) Geocode

There are now three different priority groups of address element combinations, each with their own geocoding process:

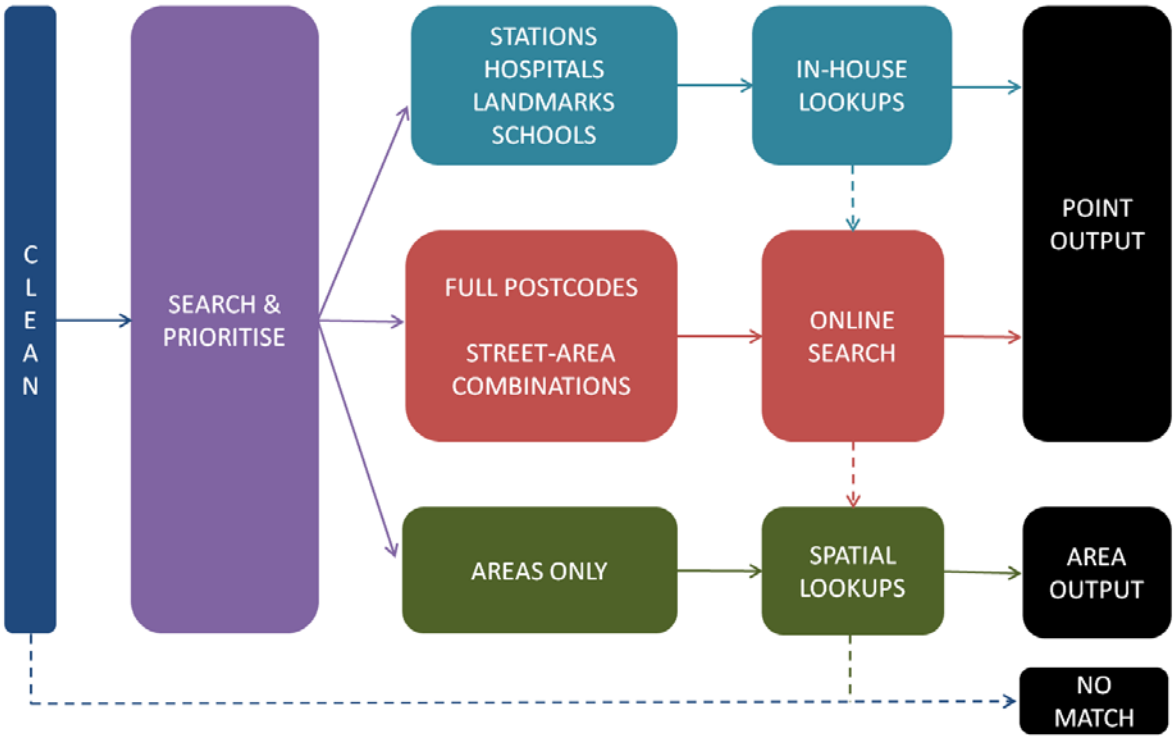
Priority 1: Hospital/Station/School/Landmark – able to geocode to a specific point on a map utilising in-house lookup tables of exact geographic co-ordinates;

Priority 2: Full postcode or anything containing a street – able to geocode to a postcode polygon centroid or street centrepoint respectively, using free online street-geocoding software;

Priority 3: Partial postcodes and Areas too large to geocode to a street – geocoded to a Local Authority boundary using in-house spatial map layers.

All of these lookups/searched can be carried out from within the FME software.

These processes are illustrated in the graphic below:



Although it is expected that the splitting and prioritising processes outlined above will ensure that records will be geocoded in the proposed ways, it is also expected that some records will not. For example, a newly-built school, a landmark not previously seen in the data, or a misspelling of a school. To cater for this the geocoding process incorporates a 'cascade' methodology where those records that cannot be geocoded at the highest accuracy level via in-house lookups are passed to the online search, and then if no match, to the spatial lookups, or output finally with no match if there is no success with any of the preceding methods.

With the ongoing aim of the process being to pass as much as possible through the in-house lookups (to ensure the highest accuracy), those records that drop

out of each of the above stages are recorded and analysed to identify learning points. If any changes can be made to the cleaning stage or in-house lookups to assist in the geocoding of these records then this is done and these records re-run through the process.

Additional processes

1. In some cases, due to duplicate streets in London, there is not enough information linked to a street-based location in a record to be able to locate to a specific location (eg. Princes Avenue, London). In these cases, all street-based records under Priority 2 above are first passed through a lookup table of London's duplicate streets and the electoral wards in which they sit. If there is no match (or a match but additional geographic information such as partial postcode or area is held) the record continues on for street matching; however if there is a match then the value of the record is split equally across the electoral wards that the duplicate street exists in. For example, there are five Princes Avenues in London, and so instead of being able to allocate it to a single location with a confidence score of 1, the score gets split and 0.2 assigned to each of the 5 electoral wards.

2. As mentioned in Stage 3(ii), 3 variations of street names are recorded to capture streets with varying numbers of words. The in-built online FME geocoder first searches based on the longest option (X Y Z street) and outputs any matches, followed by the next (Y Z street), followed by the shortest (Z street). Testing has shown this to be the most accurate method to capture the correct address. It is accepted that there is a small possibility of inaccuracy here if a shorter correct option (eg. Kent Road) is geocoded as the longer option first if the additional words prompt an 'early' geocode (eg. Old Bar, Kent Road => Old Kent Road).

3. For unknown reasons the in-built FME online geocoding tool does not parse approximately 10% of the street-area locations when searching on the separate street-area fields. For this reason the process is repeated for these 10% using the full freetext of the location rather than just the fields. It is accepted that this has an inherent risk of inaccurate results (as mentioned in the Background section of this document) and as such these 10% of records will be noted as 'Online lookup generic' as opposed to 'Online lookup specific' in the outputs.

4. Join & Spatial alignment

As a result of the previous stage, each record that contains potentially geocodable information can now be appended to one another with the below newly-assigned fields:

- the location elements used to conduct the geocoding (as per 3(iii) above)
- the method by which the record has been geocoded

- In-house lookup
- Online lookup (specific and generic)
- Duplicate Street spread
- Partial Postcode lookup
- Area lookup
- None

- a geographic level which equates to a confidence level: point (exact or street/postcode centrepoint), electoral ward, borough, or none
- a confidence score of either 1 (match to a point/area), 0 (no match) or a value in-between for duplicate street spreads.
- the relevant geographic details (easting/northing, ward code, borough code).

These fields allow users to be able to differentiate between results of varying qualities and make decisions on any filters that they wish to apply when analysing the dataset.

In order that analysis can be carried out at borough and ward level, those records that have been output with only point details are then assigned their parent wards and boroughs, with the same going for anything output at a ward level. This enables a complete 'top-down' borough approach where all low-confidence geocoded data is visible, then moving to a medium-confidence ward-level (filtering out those records that could only be geocoded at borough level), then finally to a high-confidence point-level (filtering out those records that could only be geocoded at ward level). This is illustrated in the below graphic.

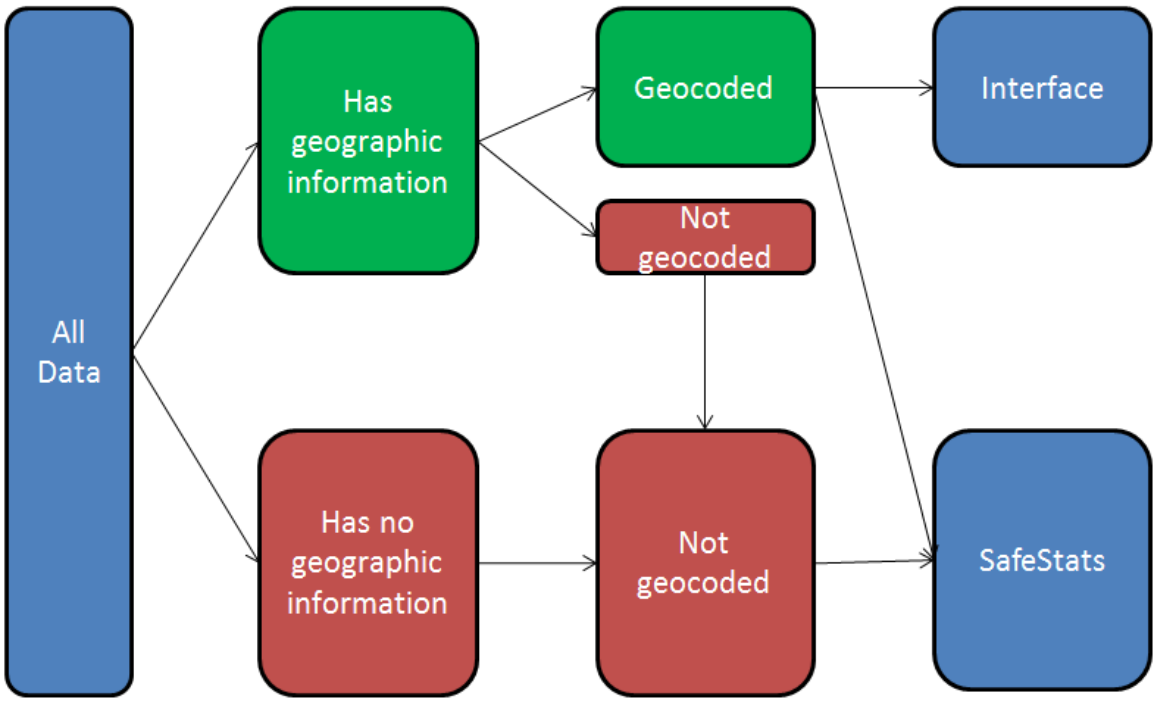
5. Disclosure

Whilst data provided to SafeStats by hospitals is non-disclosive, to ensure that no information that could potentially identify a victim of violence through their address 'slips through the net', a number of additional keyword checks are built in to the process. Naturally this only applies to point outputs, and so where any issues are identified, these are removed and ward details are then available as their most accurate geocoded location.

6. Outputs

An online interactive mapping interface has been built to display the three geographic levels of data, as well as the ability to filter on key fields, and download the data. The data is also required for hosting within the SafeStats Data repository, to be analysed alongside already hosted data from the Metropolitan Police Service, London Ambulance Service, London Fire Brigade, Transport for London and the British Transport Police.

The two output destinations however have different requirements: the interface requires only geocoded data that can be mapped, whilst SafeStats requires the full dataset (geocoded and non-geocodable). In order to achieve this, any 'No match' outputs from the geocoding output in Stage 5 above are first removed and added to the non-geocodable dataset from Stage 2 (with 'no match' replaced with 'Insufficient address information provided' to match). This now creates an updated 'Non-geocodable' dataset. What remains from the Stage 5 dataset is now suitable for the mapping interface, and is uploaded automatically by FME into cloud-based database tables. The non-geocodable dataset is then joined to the geocoded dataset and uploaded to the SafeStats Data website for access by its users. This is illustrated in the graphic below.



For further information on this process, please contact safestats@london.gov.uk