International Journal of Spatial Data Infrastructures Research, 2016, Vol.11, 73-97

An Improved Parcel-Based Approach to Bruneian Geocoded Address Database*

Nor Zetty Akhtar Abdul Hamid[1,4], Samsung Lim[2] ,Sanjeev Jha[3]

[1]University of New South Wales, zetty.hamid@student.unsw.edu.au
[2]University of New South Wales, s.lim@unsw.edu.au
[3]University of Manitoba, Winnipeg, Canada, jha.sanj@gmail.com
[4]Survey Department, Ministry of Development, Brunei Darussalam

Abstract

A new framework of Brunei's national geocoded address database is proposed in this paper. The proposed framework is based on the concept of land parcel-based geocoding and deterministic record linkage, which involves three datasets: the national address database, cadastral polygons and building centroids. The technique used in the development of the framework is an improved version of land parcel-based geocoding with no matching address components since addresses are sourced from the authorised national address database. Addresses are mapped onto the centroids of building polygons resulting in formation of geocoded address points. Cadastral polygons of land parcels act as a mediator to link the address database and the building centroids using its unique key known as 'lotnum_bc'. The proposed approach has an advantage in terms of fitting into the currently available resources. Furthermore, the proposed approach produces geocoded addresses for buildings when compared with valid addresses from the authorised address database up to the accuracy of parcel-based geocoding level. The deterministic record linkage requires validation of 'lotnum_bc' within the address database to ensure such an accuracy. It is expected that the proposed geocoded address database will become an integral part of the spatial data infrastructure of Brunei.

DOI: 10.2902/1725-0463.2016.11.art5

Keywords: geocoded address, parcel-based, deterministic record linkage, cadastral polygons

1. INTRODUCTION

For decades, location data have been a crucial component in decision making, especially for governments.Wallace et al. (2006); (Holland et al, 2009; Paull, 2012). Location information of a government's geospatial infrastructure can facilitate relationships between people, business and government (Masser et al, 2008; Williamson et al., 2006). Holland et al. (2009) stated that an effective way to represent location is to use the geocoded address reference database, which is a dataset that contains addresses with spatial references.

Ordnance Survey in the United Kingdom (UK) created a reference dataset known as 'Address Point' where all addresses from Royal Mail's Postcode Address File were georeferenced using ground survey by Ordnance Survey (Ordnance, 2010). However, a notice of product withdrawal for Address Point was announced by Ordnance Survey in October 2014 (Ordnance, 2014a) and new products known as 'AddressBase', 'AddressBase Plus' and 'AddressBase Premium' were introduced, aiming for fulfilling customers' current and future addressing needs (Ordnance, 2014b). The unique Ordnance Survey Address Point Reference (OSAPR) in Address Point is now replaced by the unique Property Reference Number (UPRN) in AddressBase products, which will be used throughout the address lifecycle. These new products include the time dimension within their datasets, and the status flags used for address status and quality in Address Point have been extended to include the address lifecycle in the AddressBase products.

On the other hand, the Public Service Mapping Agency (PSMA) in Australia has initiated the development of Geocoded National Address File (G-NAF) to centralise the national datasets for Australia. The G-NAF methodology as highlighted by Paull and Marwick (2005) was to use the address passes levels where cadastral parcels are used at the parcel level, street midpoints within the locality are used at the street level and centroids of locality at the locality level. In other words, each component in an input address can be matched with the related datasets. At the street level, for example, the street name of an address matches with the street name of the street centreline dataset within the locality, and the midpoint of the matched street segment is the estimated address location. Quality and reliability of the addresses in G-NAF are highlighted by the status flags in the dataset.

A street address database requires reliable street centrelines for the interpolation process. There are a few elements required for the interpolation process e.g., 'address distance range' i.e. the average distance between two continuous addresses, 'parity' i.e. the direction from the start node to the end node of a road

centreline, and 'offset' i.e. the distance from the exact address to the start node of a road centreline. One of the well-known street address databases is the Topologically Integrated geographic Encoding and Referencing system (TIGER) / Line developed by the United States (US) Census Bureau. The interpolation process using street centreline segments requires regular numbering patterns. Irregular address patterns and long distances from street centrelines are obvious constraints on the street centreline interpolation, which can be solved by using parcel-based geocoding, as claimed by Zandbergen (2008). However, parcel-based geocoding requires physical addresses within the land parcels (Rushton et al., 2006; Zandbergen, 2008; Edwards et al., 2014) and the reference data of this type is normally sourced from tax parcels (Edwards et al., 2014; Miranda et al., 2012). Hence, having no tax parcels or addresses within the land parcels is the main limitation of this approach. Locality addresses and post code addresses are another geocoding levels but they are not as useful as address points, street addresses and parcel addresses, especially if they are large in size.

The capability to store large data in a stable database management system (DBMS) and the availability of tools handling spatial data through Geographical Information System (GIS) have made the development and management of the geocoded address reference dataset feasible. One dataset normally aims for one type of information, and therefore a number of datasets are combined to provide a variety of useful information. The record linkage is one way of relating or combining information. One of the earliest works in record linkage was conducted by Newcombe et al. (1959), which was followed by formulation of a mathematical model by Fellegi and Sunter (1969). Research by Winkler (2006) used the record linkage process to relate people and businesses via location or address information, and also a way of cleaning data and identifying objects. In geocoding, record linkage is used for address matching. Zandbergen (2008) explained two approaches of record linkage; probabilistic and deterministic approaches. The deterministic approach assumes error free fields and exact match values, whereas the probabilistic approach has a certain degree of match accuracy. The former is more definite than the latter as the result is either accepted or rejected. GIS tools are able to support the record linkage process through either link of field attributes or spatial relationships.

A geocoded address database can be used to link spatial data with non-spatial data. Accessing, sharing and integrating spatial data are commonly carried out through the spatial data infrastructure (SDI) based on the standards and policies. In this study, a means to improve the existing approach is explored by considering the constraints of address validity and availability in land parcel-based matching for the creation of geocoded address reference dataset. A new approach aims for definite and reliable output. Brunei is chosen as the study area mainly because of two reasons: 1) Brunei requires location data for its planned Spatially Enabling Government (SEG), and 2) the existing national address database in Brunei is still

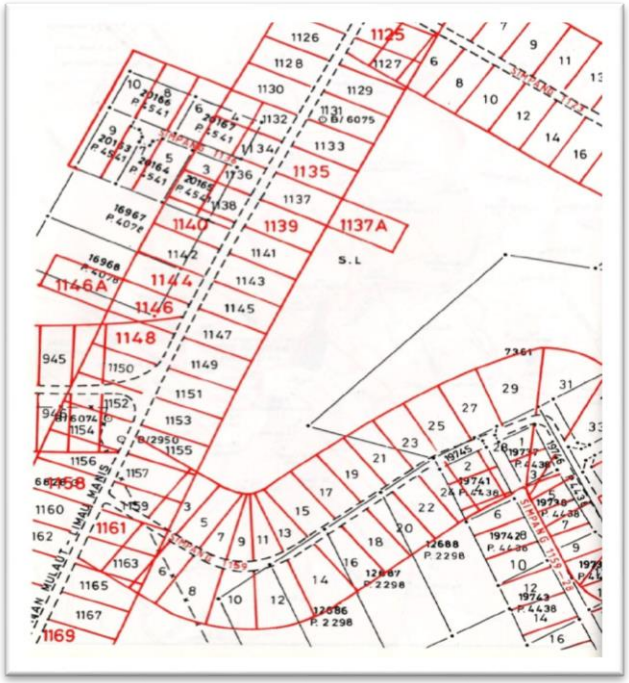
in textual format and therefore is not able to provide proper location information. With the availability of location data in the future, the Bruneian government will be able to incorporate it within its Spatial Data Infrastructure (SDI) and thus beneficial to many agencies. The overview of the existing Bruneian address database, spatial datasets and methodology for geocoded address are explained in the next section. Then the formulation of the Bruneian geocoded address reference dataset model is presented, followed by discussion and conclusion.

2. OVERVIEW OF BRUNEIAN ADDRESS DATABASE

Brunei is an independent sovereign sultanate with no local government based on the constitutional law (Bruneiresources, 2005). It is located within Borneo Island with area of 5,765 km^2. There are 55,983 addresses in Brunei, which were recorded on September 2014 by the Survey Department, the sole agency responsible for the national address database. Registered addresses are mostly residential and located within four districts. They are stored in textual format in Oracle database. Attributes of the Brunei address database include 'recno' i.e. record number, house number, street number, road name, village name, sub-district, district, postcodes, and 'lotnum_bc' i.e. lot number, applicant details, etc. Recno is the unique key of address records, lotnum_bc is the unique key of land parcels, and the record details include name, address, etc. A house number is not a null attribute. Brunei's addresses can have either a street number or a road name or both. Based on the house and junction numbering guidelines for Brunei (2014), the address number pattern depends on the side of the road or street segment from the start node. In this paper, street and junction refer to the same entity. A start node normally falls within a junction point where the road or street segment starts. Addresses with odd numbers will be on the left side and even numbers will be on the right side. Currently the address numbers are estimated by Survey Department using the 'numbering block', which is based on road and street segments, and physically drawn on cadastral maps. A cadastral map is a map containing land parcels and roads, and is printed by grid numbers. The numbering block contains address numbers that are reserved for houses or junctions within the block. A land parcel, which is represented by a cadastral polygon in this paper refers to an authoritative land boundary collected using survey methods. Each block with one address number is 20 m (parallel) by 40 m (perpendicular) to the street segment (Brunei, 2014). Figure 1 illustrates the estimation of address numbers based on the aforementioned criteria. There is also a possibility of including alphabets at the end of an address number if there are more than one building using an access road that does not have a street number. For example, a house number '1137A' in the middle of Figure 1. A typical housing type in Brunei is a single address for one building. However, there exist terrace houses and flats, which have multiple addresses within a building. These residential types tend to grow but their numbers are still small, compared to single-address houses. The national housing programme conducted by the government has recently included

these multiple-address residential types in order to resolve the housing demand and supply for citizens (Brunei, 2013). One advantage of the existing structure of the address database is the 'lotnum_bc' attribute. Even though the addresses are not given in the spatial format, this attribute can be used to link the addresses with the land parcels. However, lotnum_bc is unfortunately not a mandatory field, hence users may overlook the importance of this attribute.

Figure 1: Estimation of Addresses Using a Numbering Block on Top of Cadastral Map. Sourced from (Brunei, 2014).



As stated above, Brunei's national addresses are available only in the textual format. Inconsistent developments between the government agencies are among some reasons of not having a national geocoded address database yet. However, awareness of the benefit of spatial information has been phased in over the last decade.

Address points normally have a high positional accuracy, compared to street addresses. Parcel-based addresses, on the other hand, can resolve the limitation of street addresses in the case of an irregular address pattern and a long distance of the address from the street centreline. Thus, in this study, address points and parcel-based addresses are tested with respect to the Bruneian addresses. The

main objective is to produce an address point geocoded address reference dataset for Brunei. It is the right time for Brunei to create the location data before the datasets become more complicated and complex. In this paper, the term Geocoded Address Point will be used instead of Geocoded Address Reference Dataset for the sake of simplicity.

The total number of Bruneian national address records is less than sixty thousands. The Brunei Government is trying to avoid the manual approach using Global Positioning System (GPS) as used by the UK Ordnance Survey for the creation of their geocoded address database. In a preliminary analysis, it is observed that some components in addresses such as street names and road names are not consistent among datasets. Thus, the approach used for G-NAF is also impractical for Brunei. Not having topology and address range attributes in the Brunei road centrelines may cause difficulties in adopting a similar approach to TIGER, too.

3. DATASETS AND APPROACHES FOR CREATING GEOCODED ADDRESS POINT

An address complements the land administration and management as it helps people identify locations easily. A road name and a street number within an address are physically labelled, whereas cadastral information is not something that visually stands out. However, a normal address will not be able to provide location digitally without any spatial reference. As stated previously, both Address Point in UK and G-NAF in Australia are using mapping data and Digital Cadastral Database (DCDB), respectively, for the source of their geocoded address points. Golberg (2008) claimed that building footprints gathered from surveys can be the most accurate data source. However, Kalantari et al. (2008) argued that the unstable address of a building can create a problem in the future especially with a legal perspective in land administration activities and that not all buildings have an address. Spatially-referenced legal-property objects share spatial identifiers to maintain their relationships among each other (Kalantari et al., 2008). From these reviews and comparing to Brunei existing resources, five datasets have been identified to be useful in this study:

- National Address Database (NAD)
- Cadastral land parcel that is also referred to cadastral polygon (CP)
- Building polygon (BP) and its centroid (BC)
- Road Centerline (RC)
- Admin Boundary Polygon (ABP)

Preliminary analysis on the datasets shows that record linkage and spatial relationship are suitable techniques in creating Bruneian geocoded address point. It is the authors' aim to use 'recno' as the unique key of addresses, stored in a correct building polygon and its centroid point. A cadastral polygon acts as a mediator between the two datasets through its unique key lotnum_bc, which exists in the address database and can be spatially integrated with building polygons/building centroids. The record linkage used in this study is deterministic rather than probabilistic as there are only two outputs to the value of lotnum_bc: correct or wrong. The common key, which refers to lotnum_bc must be present in the three datasets: the national address database, cadastral polygons and building polygons/centroids. This approach is regarded as an extension to parcel-based geocoding, but instead of using address from tax parcels, they are extracted from the national address database, which is a more reliable source. The final address points will use building centroids rather than parcel centroids. Datasets are categorised into two types: source data and reference data. Source data contributes to the newly created geocoded address points, whereas reference data are used mainly for validating results.

G-NAF in Australia and Address Point in UK used additional attributes to highlight the status of an address. The assigned values must have a clear description and a set of values will cover the whole scenario. The status flag 'link_flag' is created in the address database to indicate the integration status of an address to the other two datasets. The possible values for link_flag are 0 (no integration at all), 1 (integrate with cadastral database only) and 2 (integrate with both cadastral and building). Link_flag of value 2 means that the address can be geo-referenced into an address point and can be used in Brunei geocoded address points. However, having more than one address within a cadastral polygon will provide uncertainties when linking addresses to building centroids. As a result, link_flag value will stay as 1 (but not 2), unless the address location can be confirmed. The link_flag value -1 as explained in Abdul Hamid et al. (2015) for 'unresolved' is handled differently in this study. This scenario of having a set of addresses within a cadastral polygon is called a multiple-address point. Figure 2 illustrates the process of using record linkage for Bruneian geocoded address points with the following steps that are based on the framework model by Abdul Hamid et al. (2015).

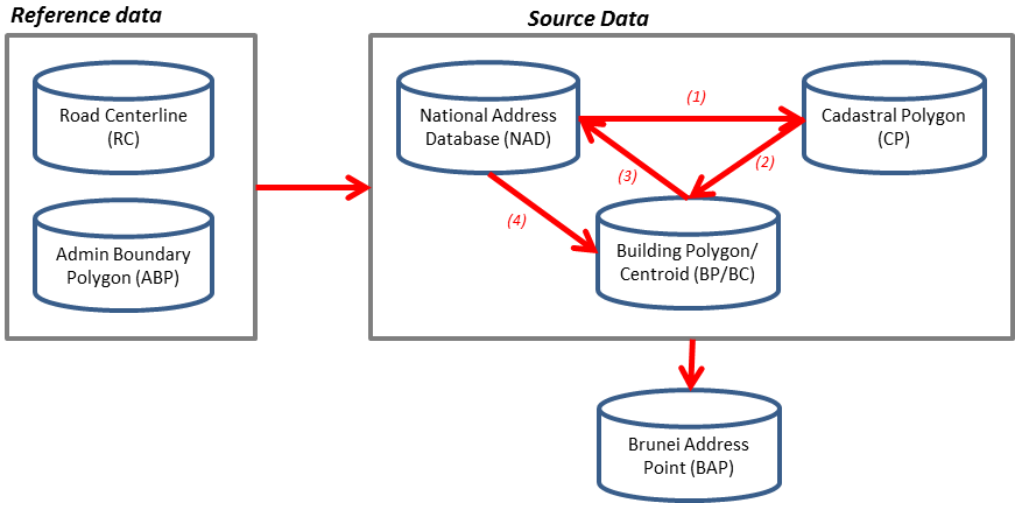
Step 1: To link the address database and cadastral polygons via 'lotnum_bc'.

Step 2: To link cadastral polygons and building polygons/centroids spatially and to update the value of 'lotnum_bc' in building polygons/centroids.

Step 3: To verify the link between building polygons/ centroids and the address database, and to update 'link_flag' values.

Step 4: Finally, to update the 'recno' attribute in building polygons/centroids. Brunei address point (BAP), a new dataset, is a subset of Building polygons/centroids (BP/BC) with link_flag equal to 2.

Figure 2: Deterministic Record Linkage for Brunei Geocoded Address Point (Based on Abdul Hamid et al, 2015).



Further to the above approach, three stages are conducted to refine the Brunei geocoded address point framework.

Stage 1

The main purpose of this stage is to test the process outlined in Figure 2. The result is analysed for the initial assessment of the approach for geo-referencing addresses. Some adjustments are made to the data and the approach as below:

- A frequency tool is used for both the address database and building polygons based on grouping of lotnum_bc. Results from both datasets are then integrated and the number of addresses and buildings for each cadastral polygon will be known. This approach will be able to identify the relationship between the two datasets that will affect the link_flag value;

- Any addition or extension of a building will share the same address as the main building. They may be captured within the same polygon or as a different polygon from the main building. The former will not affect the analysis process, but the latter will produce an invalid relationship between the address and the building. Based on random checking of the sample area, building polygons with size < 100 m2 are assumed not to have their own address and are excluded

from this analysis. However, they will be retrieved back after the relationship has been obtained;

- A building centroid is preferred than a building polygon because of the limitation of the spatial join tool that only considers the condition of 'completely within'.

After the above adjustments are made, the summary results are provided in Table 1. The table shows that 17,073 addresses (about 30.5%) are successfully linked with buildings based on the ideal one-to-one relationship between one address and one building within one cadastral polygon. The most obvious errors fall within the zero building relationship where 30,865 addresses (about 55%) did not have any relationship with buildings and land parcels. The main problems are inaccurate lotnum_bc in the address database and the outdated building data. Further investigation on the zero building relationship indicates that 43% addresses (13,213) within this group have missing lotnum_bc. Others are caused by typo errors during the data input, out-of-date lotnum_bc in the address database (because the survey process updated the cadastral polygons and the unique key lotnum_bc, but this has not been reflected to the lotnum_bc in the address database), out-of-date addresses (e.g. some buildings were demolished) and new addresses (e.g. buildings not yet built or updated). Other relationships in Table 1 refer to different types of multiple addresses. Link_flag of each relationship is also reflected in the table.

Table 1: Relationship between Buildings and Address within One Cadastral Polygon.

Relationship	Building	Address	No. of Records	Select Statement's condition	Findings	Link_flag
Zero building	0	>0	30865	"freqnbld" =0	Building or lotnum_bc in address data not accurate or updated.	0
1 to 1	1	1	17073	"freqnbld" =1 AND "freqnaddr"=1	Ideal.	2
1 to many	1	>1	3167	"freqnbld"=1 AND "freqnaddr">1	Mostly terrace and flat	1
Many to 1	>1	1	1938	"freqnbld">1 AND "freqnaddr"=1	Mostly non-residential such as school.	1
Many to many (consistent)	>1 (x)	>1 (x)	1420	"freqnbld" = "freqnaddr" AND "freqnbld" <>0 AND "freqnbld" <>1	Ideal with the number of addresses is equal to the number of buildings. But need	1

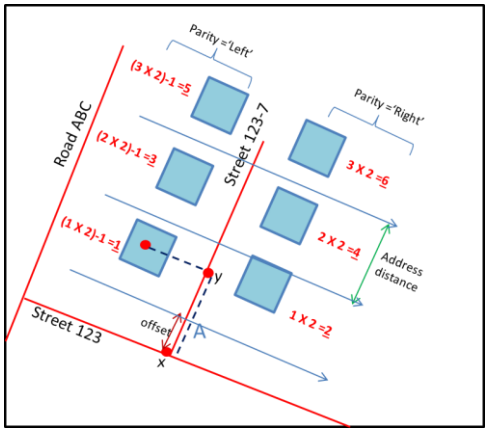
					correct mapping between both datasets.	
Many to many (inconsistent)	>1 (x)	>1 (y)	1520	"freqnbld" <> "freqnaddr" AND "freqnbld" <>0 AND "freqnbld" <>1 AND "freqnaddr" <>1	Too complicated. Need to solve case by case.	1
TOTAL			**55983**			

Our results show the importance of ensuring accurate values of lotnum_bc in the address database as a link key to cadastral polygons and building polygons/centroids. Making it as a mandatory attribute is necessary to avoid users from missing it again, but the existing errors need to be resolved first. Based on the results at this stage, link_flag attribute values can be extended on the basis of the three initial values 0, 1 and 2. The status of an address such as an expired address and a new address can have their own link_flag value.

Stage 2

In the previous stage, it was observed that there are many types of multiple addresses with link_flag equal to 1. In this stage, an alternative approach is proposed to use road centrelines in the estimation of address points for multiple addresses in Table 1. The goal is to check the reliability of the road centrelines when the interpolation process is complete. It is worth pointing out that the results obtained from this approach can only be assessed manually rather than automatically due to the unavailability of ground truth data as a reference. The existing road centrelines do not have any network topology or address range attributes, mainly because it was created for the purpose of cartography. Its attribute includes road_name, f_code (feature code) and other spatial attributes. The feature code, f_code, specifies the road type through code numbers. By considering elements in estimating the address number such as 'offset', 'address distance' and 'parity' as shown in Figure 3, Equations (1) and (2) are derived and used within the interpolation process for the address number estimation. In many studies, offset values show different impact on positional accuracy (Zandbergen, 2009; Cayo and Talbot, 2003; Ratcliffe, 2001), i.e. selection of value is important.

Figure 3: Logical Concept of Address Numbering Based on Location of the House or Building from the Road Centreline. Road 'Offset' (Distance from Start Node of the Road Centreline), Address Distance (Average Distance between Two Continuous Addresses) and Parity (Position of Address Based On Road Centreline).



Left: $$\text{Number} = (2 \times \left\{\text{Upper}\left(\frac{A-offset}{address\ distance}\right)\right\}) - 1 \quad (1)$$

Right: $$\text{Number} = (2 \times \left\{\text{Upper}\left(\frac{A-offset}{address\ distance}\right)\right\}) \quad (2)$$

This approach resolves some of the multiple address issues provided that there are no other pertaining issues as highlighted in Table 2. One example is shown in Figure 4 for lotnum_bc '100MD', where two addresses out of six are incorrectly numbered. Two smaller polygons are found to be out of the buffer region for parity as highlighted in Item 6 of Table 2. A number of issues with the road centrelines, mainly the way the data were collected and updated, was observed. Fixing and adjusting the road centrelines to fit the purpose of the address number estimation can be time consuming and costly. Both processes require validation for all road segments, which might be similar to collecting a new dataset. Some issues highlighted in Table 2 require references to the numbering block.

Figure 4: Address Estimation for Lotnum_Bc '100MD'. Four Addresses ('3', '4', '5' And '6') were Correctly Estimated. However, four Buildings' Centroids Have Incorrect Addresses ('0') due to Constraints on Corner Buildings and not within Parity Buffer.

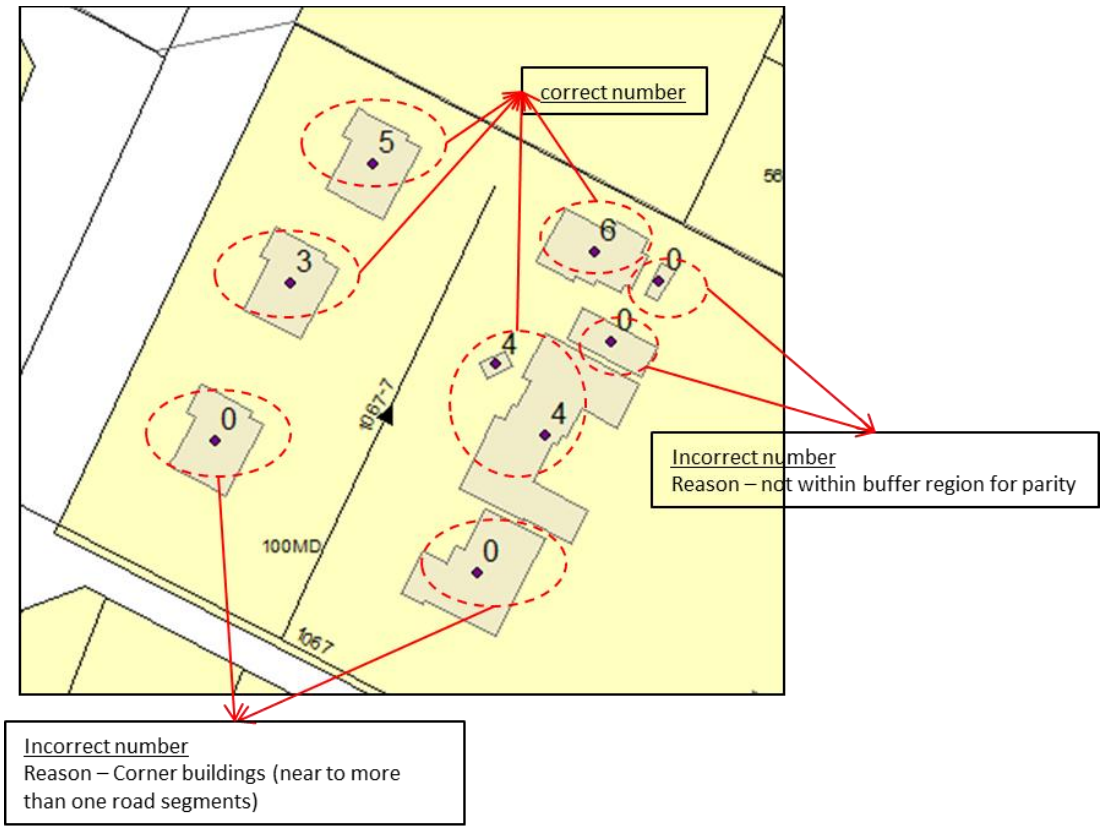


Table 2: List of Issues Using Existing Road Centrelines.

No	Issues	Description	Tentative Solution and limitation
1	Many segments for each road or street.	Many segments mean many start nodes for each road or street. Thus, address numbering will be incorrect.	To join segments using 'Unsplit' tool in ArcGIS. Require validation because some segment did not join perfectly.
2	Short road or street segment.	There are many scenarios with short segment: i. A typical access road, which does not contribute to address	For scenarios i. to iv., there are three solutions: a. Remove segments for scenarios i. & iii.

		estimation or pattern as in Figure 5. ii. An access road that cause a non-straight address number pattern as in Figure 6. Address numbers were results from stage 1. iii. As a merger of two roads or streets and it does not contribute to address estimation. iv. As a merger of two roads or streets and its attribute value is used as part of the address such as Line 2 in Figure 7. However, address interpolation is using Line 3.	b. Remove segment for scenario iv. but retain the value. c. Scenario ii. is complicated if to be automated, as it needs to be joined with main segment and parity will depends on the main segment. 'Remove' tool in ArcGIS is used to remove segments. Identifying the scenario for the short segment in an automated way is not a straight forward process.
3	Road segment has wrong direction.	Wrong direction affects the parity, which is one element in estimating address number.	Display segments' direction on the map. Identify segment with wrong direction and use 'Flip' tools. Identifying segment with wrong direction will takes time. The unreliable feature code of road centreline limits the automation process.
4	Road segment's direction cannot be predicted	Road segments have both ends intersect with same road type, such as both ends joined to major roads. Thus it is difficult to assume the direction.	Segments are selected when both ends intersect with segments of same road type based on feature code value of the segment. However, it requires reference to the 'numbering block' for the correct direction or address numbers.
5	Inaccurate Road's name and street's number.	Road name and street number are part of address. The inaccuracy may be based on: i. Typo error during data collection. ii. The 'road name' attribute in road centreline is inconsistent with 'road name' and 'street number' attributes in address database.	Labelling the road centreline using 'road name' attribute is necessary for data validation. Inaccurate data in items i, ii and iii in need to be corrected or updated. For item ii, inconsistency mainly caused by inclusion of term 'JALAN' and 'SPG' in road centreline, which refers to road and street respectively. Those

		iii. The road name has been formally changed or the street has been upgraded to road. In both cases, data are not updated in address database and/or road centreline.	terms were used in address database as they are stored in different attributes. Thus it is recommended to have a new attribute 'Road_type' in road centreline to store those terms. For item iii, renaming of road name or street number need to be reflected in both address database and road centreline.
6	No parity attribute in building polygon.	Parity indicates whether building has odd or even number.	Buffering at certain distance individually from both sides of road segment. Intersection of buildings with buffer region will give parity value. Buffer distance is subjective. It may depend on the road type, road width or other factors. If it is shorter more buildings will be excluded. And if it is longer there will be more corner building (as in item 7).
7	Corner building	Corner buildings are nearer to more than one road, causing it to be difficult to choose, which road segment to be used for interpolation of an address.	Identifying corner buildings is easy by using 'intersection' tool in ArcGIS for buffer process in item 6. But selection of road segment needs reference to 'numbering block'.
8	Alphabets in address number	Calculation for even or odd numbers will be affected	Need to remove the alphabets to new attributes, prefix and suffix.
9	Inconsistent feature code.	Feature code is important in supporting resolving some issues above. Assumptions on the inconsistency are users have different judgement on road type and feature code has not been updated with any road change or upgrade.	Feature code need to be corrected. Using 'label' to show the feature code will assist checkers in correcting errors.

Figure 5: Shorter Road Segments that Shows a Typical Access Roads which will not Be Used for Address Estimation.

Figure 6: A Sample of Access Roads that Cause a Non-Straight Address Numbers. for Instance the Bottom Arrow Line Highlight Access Road that Contribute to Addresses '1', '3', '5' and '7' That are not in Straight Line. Addresses are obtained from Stage 1 of this Study and Estimation Using Road Interpolation is not possible.

Figure 7: Sample of Access Road (Line 2) that Act as Merger to Two Road Segments (Line 1 and Line 3). Its Attribute Value 'Spg 1067-12' is used for All Addresses within the Dotted Block. Addresses inside the Dotted Block are obtained from Stage 1 and Visually They Are Based on Line 3.

Stage 3

Based on Stage 2 above, the manual approach is required to resolve some complicated issues. Findings from Stages 1 and 2 provide enough information for refining the framework illustrated in Figure 2. The adjustment is shown comprehensively in the Bruneian geocoded address point model in Figure 8, where both automated and manual record linkages are used. In the manual approach, recno is manually input into building centroids. Uncertain locations need references to the 'numbering block'. Each address can be manually linked within 30 seconds to 3 minutes depending on its complexity. Link_flag values are extended based on the issues detected from the first stage. Values 0 to 9 are reserved for various possible scenarios whereas values 10 and above are reserved for errors or issues as described below:

i. 0 – address has no linkage with cadastral and building (possible data errors)

ii. 1 – address has linkage with cadastral only

iii. 2 – address has linkage with both cadastral and building

iv. 3 – address expired (building demolished)

v. 4 – address has been given, but building is not yet finished

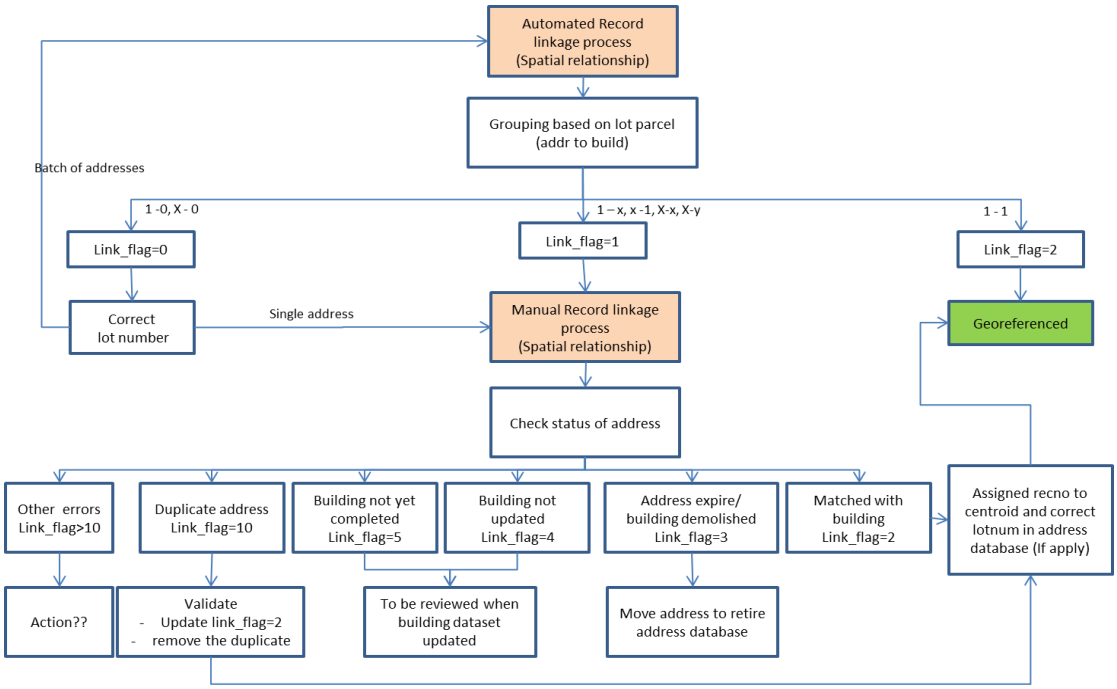
vi. 5 – building is not yet updated

vii. 6 to 9 – reserved for new unidentified scenarios

viii. 10 – possible duplicate address

ix. Above 10 – reserved for new unidentified errors or issues

Figure 8: Brunei Geocoded Address Point Process Flow.



The 'georeferenced' box in Figure 8 contains a list of addresses that have been successfully geo-referenced as address points with link_flag equal to 2. The figure also shows actions for different link_flag values.

4. DISCUSSION

The proposed framework in Figure 8 has improved the existing technique on geo-referencing addresses based on the concept of parcel-based geocoding, deterministic record linkage and spatial relationship. There are only three source datasets required to create the geocoded address points by integrating datasets via common keys 'lotnum_bc': the national address database, cadastral polygons and building centroids. Cadastral polygons act as a mediator to link the addresses with the building centroids. The unique key 'recno' from the address database is

copied into the building centroid database for the successful georeferencing. The use of status flag 'link_flag' to indicate the status of the address to some extent will assist users to identify addresses with different attention level. Advantages of this approach are as follows:

- Address availability and validity are guaranteed as addresses are sourced from the authorised address database;

- The definite outcome from the deterministic record linkage is more reliable, especially without any ground truth data;

- Less datasets are required for address matching as there is no need to do matching for address components; and

- Status flag values provide different types of address status.

However, there are also constraints as below:

- The existing 'lotnum_bc' in the address database is not a mandatory attribute and most errors are contributed by missing 'lotnum_bc';

- Out-of-date addresses and building data hinder the success of the integration process;

- Multiple addresses i.e. more than one address or/and more than one building within one land parcel or cadastral polygon, may complicate the allocation of address numbers; and

- Parameter values for the sizes of buildings used to indicate whether it is a main building or just extension of main building will reflect frequency relationship of address and building.

The first constraint can be resolved by converting the attribute to mandatory after all missing values are resolved. The second constraint can be resolved by using updated and validated datasets. The third constraint can be resolved by either using the street centreline interpolation or preferably reference to the numbering block. The fourth constraint can be resolved by ensuring each building polygon has an attribute indicating the main building or extension. If all constraints are resolved, the lower success rate (about 30%) can be significantly improved.

Comparisons between this approach and other existing known approaches are summarised in Table 3 including their accuracies and challenges. Selection of method by researchers as stated by McElroy et al. (2003) must be through thorough understanding of the geocoding process to geocode the study area and

assess the quality of the result. This may refers to the availability and reliability of resources, costs, benefits and precisions required for the results. In the case of Brunei, the method record linkage from G-NAF for address matching is adapted in the first link. The second link uses spatial relationship tools in GIS. However, the first link is not using the common address components as in G-NAF but uses the land parcel unique key and the address unique key. The result provided accuracy and precision at buildings' centroid. In having one address in one building and in one land parcel (1-1-1 relationship), the method will work in a situation below that can be a limitation to other approaches, provided that the land parcel unique key is one of the 'mandatory' attributes in the address dataset:

- Inconsistent values in address components such as street names;
- Irregular address patterns; or
- Any address format.

The method is suitable for and recommended to places or areas with the majority of single family detached houses in one land parcel. Both G-NAF and TIGER will have a constraint with the first situation above, and TIGER will not work with the second situation. Even though the study area is small, it is expected to work in a larger area that complies with the above conditions. In this study, we also observed that the manual approach is preferred for Brunei as it is expected to be more economical than new data captures or fixing unreliable datasets. However, there will be no manual intervention for new addresses once the land parcel unique key in the address dataset is made mandatory. In this technology era with the availability of GPS, a cheaper version of data capture can be carried out using the Volunteered Geographic Information (VGI). However, the major issues are accuracy and validation (Behr, 2010).

The new Bruneian geocoded address point database is proposed to be included as a national dataset based on its potential to spatially enable information, which can be shared within agencies for their improved decision making via its spatial data infrastructure. The integrity of the datasets needs to be efficiently managed to ensure their consistencies. Agunbiade et al. (2014) pointed out that a silo effect is one contributor to uncertainties in inter-agency integration. Maintaining the integrity of the common key of address and land parcel within agencies will be the next step of this study where existing land administration processes, policies and guidelines will be reviewed.

Table 3: Evaluation of the Proposed Geocode Address Methodology and Existing Methodologies.

Geocoded Address	Approach	Accuracies	Challenges	Discussion
Address Point (UK)	GPS capture	High accuracy based on manual GPS capture	Costly and time consuming	Requires constant maintenance as it requires cross check for new addresses.
TIGER (USA)	Linear interpolation of address numbers within the address range of a street centreline	Medium accuracy depending on reliable street centrelines and other parameters such as offset	Requires reliable street centrelines and address range attributes Will not works with Irregular address patterns Data updates	Offset values – small different values can give big difference in accuracy TIGER street maps are free despite the Infrequent updates (McElroy et al., 2003) Ratcliffe (2001) has conducted assessment on using TIGER type geocoding process and found more than 50% addresses fall on land parcels of different property for 20k addresses in Sydney, when assess to cadastral and census areal units.
G-NAF (Australia)	Multiple address matching techniques using record linkage and geospatial analysis (Paull, 2003; Paull and Marwick, 2005)	High accuracy depending on the address matching output	Dependent on consistent address values	G-NAF is recently openly available by the Australian Government (Australia, 2016), which can encourage more creativities
Bruneian geocoded address database	Link of textual address to cadastral and	High accuracy (building centroid) depending on	Limited to 1-1-1 relationship to address, land parcel	The current study area does not have ground truth data. Gatrell (1989) stated the needs of benchmarking data. However, in that

	building polygons. (Deterministic record linkage and spatial relationship tools)	reliability of common attribute key	and building Making land parcel unique key as mandatory attribute in address database	analysis, there were issues such as using the digitized addresses supplied by Pinpoint Analysis Ltd in assessing the method as data were not freely available and used for commercial purposes. The question pointed out was what type of data is worth digitization.

5. CONCLUSION

We proposed a framework for the Bruneian geocoded address point database, which will be used as location data for the Brunei government to be spatially enabled. The proposed framework was developed based on the concept of parcel-based geocoding where the address of a land parcel is used in the address matching process. However, unlike the typical parcel-based approach that uses addresses stored in the land parcel database such as a tax parcel database, this approach uses addresses from the textual-format national address database. These addresses were linked to the land parcels and cadastral polygons via their common key attribute 'lotnum_bc' that is the unique key for land parcels. The address component matching process in the new approach took less time, compared to what is normally practised by the traditional parcel-based approach. In this study, only one attribute from the address database is used to link addresses and land parcels, and then link to building centroids. Basically, land parcels or cadastral polygons act only as a mediator for the address database and building centroids. The integration of the national address database, cadastral polygons and building centroids through the common key 'lotnum_bc' provided highly reliable geocoded address points due to the deterministic record linkage. The results of the proposed deterministic record linkage are promising, however, there are constraints of using this approach. Complexity in multiple addresses that have more than one address and/or more than one building within one land parcel, has to be manually dealt with the existing 'numbering block', which was used by the agency when providing addresses to applicants. This study provided an alternative way to resolve this issue by using the interpolation of street centrelines, which worked for some cases, however, the results can only be evaluated visually based on references to the 'numbering block'. The other constraint was the missing 'lotnum_bc' values in the address database that affect the integration process.

As a conclusion, this approach is recommended for an address database that has a mandatory unique key of land parcels as one of its attributes. This approach is particularly suitable for single family detached houses in one land parcel, which has a 1-1-1 relationship between addresses, building polygons and land parcel

polygons. As for the Bruneian address database, it is recommended that this attribute should become mandatory and be frequently validated. Future research will cover the analysis of the inter-process integration to ensure the consistent attribute values of land parcels used within the land information management.

ACKNOWLEDGEMENTS

The authors would like to thank Brunei's Survey Department for their support of this research and provision of the datasets. The views and conclusions expressed in this paper are those of the authors and not necessarily representing any organizations or any person connected with them.

6. REFERENCES

Abdul Hamid, N. Z. A.,Lim, S. & Jha, S. (2015). Quality Control for the Development of the Bruneian Geocoded Address Database. *URISA Journal,* 27, 1, 47-56.

Agunbiade, M.,Rajabifard, A. & Bennett, R. (2014). Land administration for housing production: analysis of need for interagency integration. *Survey Review,* 46, 334, 66-75.

Australia, G. (2016). *PSMA Geocoded National Address File (G-NAF)* [Online]. data.gov.au: Australian Government, at https://data.gov.au/dataset/geocoded-national-address-file-g-naf [accessed 28 June 2016].

Behr, F.-J. (2010). Geocoding: Fundamentals, Techniques, Commercial and Open Services. *AGSE 2010*, 111-122.

Brunei, G. (2013). National Housing Programme in Brunei Darussalam: 'Homes for the Nation'. *Promotion of Good Governance: Towards Fulfilling People's Asiprations and Welfare.*

Bruneisources (2014). *House and Junction Numbering Concept - Negara Brunei Darussalam*. Surveyor General, Ministry of Development, Negara Brunei Darussalam. Available at Survey Department Office.

Bruneire, S. D. (2005). *Introduction to Brunei Darussalam (Brunei the Abode of Peace)* [Online]. The Brunei resources website, at http://www.bruneiresources.com/bruneibackground.html [accessed 19 June 2015].

Cayo, M. R. & Talbot, T. O. (2003). Positional error in automated geocoding of residential addresses. *International journal of health geographics,* 2, 1, 10.

Edwards, S. E.,Strauss, B. & Miranda, M. L. (2014). Geocoding Large Population-level Administrative Datasets at Highly Resolved Spatial Scales. *Transactions in GIS,* 18, 4, 586-603.

Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association,* 64, 328, 1183-1210.

Gatrell, A. C. (1989). On the spatial representation and accuracy of address-based data in the United Kingdom. *International Journal of Geographical Information System,* 3, 4, 335-348.

Golberg, D. W. (2008). *A Geocoding Best Practices Guide* [Online]. The North American Association of Central Cancer Registries, at http://www.naaccr.org/LinkClick.aspx?fileticket=ZKekM8k_IQ0%3D&tabid=239&mid=699 [accessed 19 June 2016].

Holland, P.,Rajabifard, A. & Williamson, I. (2009). *Understanding spatial enablement of government, Proceedings of the Surveying and Spatial Sciences Institute Biennial International Conference, Adelaide 2009, Surveying and Spatial Sciences Institute, 28 September-2 October 2009* [Online]. at http://www.csdila.unimelb.edu.au/publication/conferences/understanding_spatial_enablement_of_government.pdf [accessed 24 June 2016].

Kalantari, M.,Rajabifard, A.,Wallace, J. & Williamson, I. (2008). Spatially referenced legal property objects. *Land Use Policy,* 25, 2, 173-181.

Masser, I.,Rajabifard, A. & Williamson, I. P. (2008). Spatially enabling government through SDI Implementation. *Internationl Journal of Geographical Information Science,* 22, 1, 5-20.

Mcelroy, J. A.,Remington, P. L.,Trentham-Dietz, A.,Robert, S. A. & Newcomb, P. A. (2003). Geocoding addresses from a large population-based study: lessons learned. *Epidemiology,* 14, 4, 399-407.

Miranda, M. L.,Edwards, S. E.,Anthopolos, R.,Dolinsky, D. H. & Kemper, A. R. (2012). The built environment and childhood obesity in Durham, North Carolina. *Clinical pediatrics,* 51, 8, 750-758.

Newcombe, H. B.,Kennedy, J. M.,Axford, S. & James, A. P. (1959). Automatic Linkage of Vital Records Computers can be used to extract" follow-up"

statistics of families from files of routine records. *Science,* 130, 3381, 954-959.

Ordnance, S. (2010). *Address-Point User Guide and technical specification* [Online]. at https://www.ordnancesurvey.co.uk/docs/user-guides/address-point-user-guide.pdf [accessed 25 June 2014].

Ordnance, S. (2014a). *Address-point notice of product withdrawal* [Online]. at http://www.ordnancesurvey.co.uk/docs/legal-notices/address-point-withdrawal-notice.pdf [accessed 24 June 2015].

Ordnance, S. (2014b). *Q&A in respect of the withdrawal of Address-Point* [Online]. Ordnance Survey, at http://www.ordnancesurvey.co.uk/docs/legal-notices/legacy-addressing-qa.pdf [accessed 20 June 2015].

Paull, D. (2003). *A geocoded national address file for Australia: The G-NAF what, why, who and when* [Online]. at https://docbox.etsi.org/STF/Archive/STF321_TISPAN3_EC_Emergency_Call_Location/Public/Library/Australia/G-NAF_What_Why_Who_When[1].pdf [accessed 10 June 2015].

Paull, D. (2012). Spatial Enablement: Including 'Location' in Your Thinking, Problem Solving and Decision Making. *Global Geospatial Conference 2012 (GSDI 13).* Quebec City, Canada.

Paull, D. & Marwick, B. (2005). Understanding G-NAF. *Proceedings of SSC 2005, (Spatial Intelligence, Innovation and Praxis), Spatial Sciences Institute.* Melbourne, September 2005.

Ratcliffe, J. H. (2001). On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographical Information Science,* 15, 5, 473-485.

Rushton, G.,Armstrong, M. P.,Gittler, J.,Greene, B. R.,Pavlik, C. E.,West, M. M. & Zimmerman, D. L. (2006). Geocoding in cancer research: a review. *American journal of preventive medicine,* 30, 2, 16-24.

Wallace, J.,Williamson, I.,Rajabifard, A. & Bennett, R. (2006). Spatial information opportunities for Government. *Journal of Spatial Science,* 51, 1, 79-99.

Williamson, I.,Wallace, J. & Rajabifard, A. (2006). Spatially enabling governments: A new vision for spatial information. 17th UNRCC-AP Conference and 12th Meeting of the PCGIAP, Bangkok, Thailand. 18-22 September 2006.

Winkler, W. E. (2006). Overview of record linkage and current research directions. Technical report. *Bureau of the Census - Research Report Series.*

Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems,* 32, 3, 214-232.

Zandbergen, P. A. (2009). Geocoding quality and implications for spatial analysis. *Geography Compass,* 3, 2, 647-680.