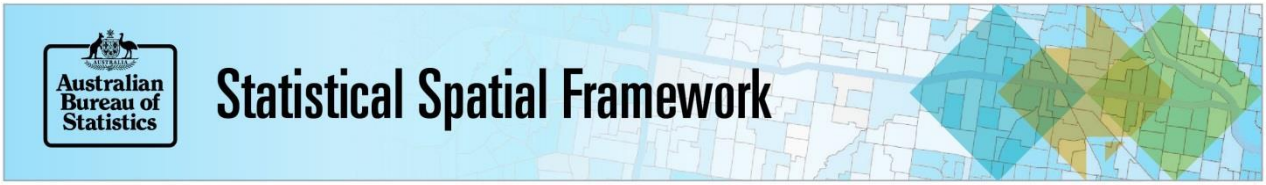


SSF Guidance Material – Geocoding Unit Record Data Using Address and Location

The Statistical Spatial Framework (SSF)[1] identifies the importance of linking socio-economic data to a location to enable that data to be used in regional analysis and reporting. To achieve this, the SSF specifies that each unit record in socio-economic datasets be linked to a location through a set of *geocodes* (defined below). This geocode information can then be used to aggregate (or combine) the unit record data for larger regions to provide summary statistics for analysis and reporting.

The SSF recommends that, ideally, the set of geocodes stored for unit records in socio-economic datasets should include both a location coordinate and an Australian Statistical Geography Standard (ASGS)[2] Mesh Block code. The SSF recognises that in some instances the location information in the dataset may not permit allocation of a location coordinate or Mesh Block and so the unit records will need to be linked to a larger region.

Purpose

There are many different ways of geocoding information; this paper covers three of the main options for geocoding unit record data in socio-economic datasets. The paper describes the basic elements and processes applied when implementing these three options, and provides references to resources associated with them.

Options for geocoding

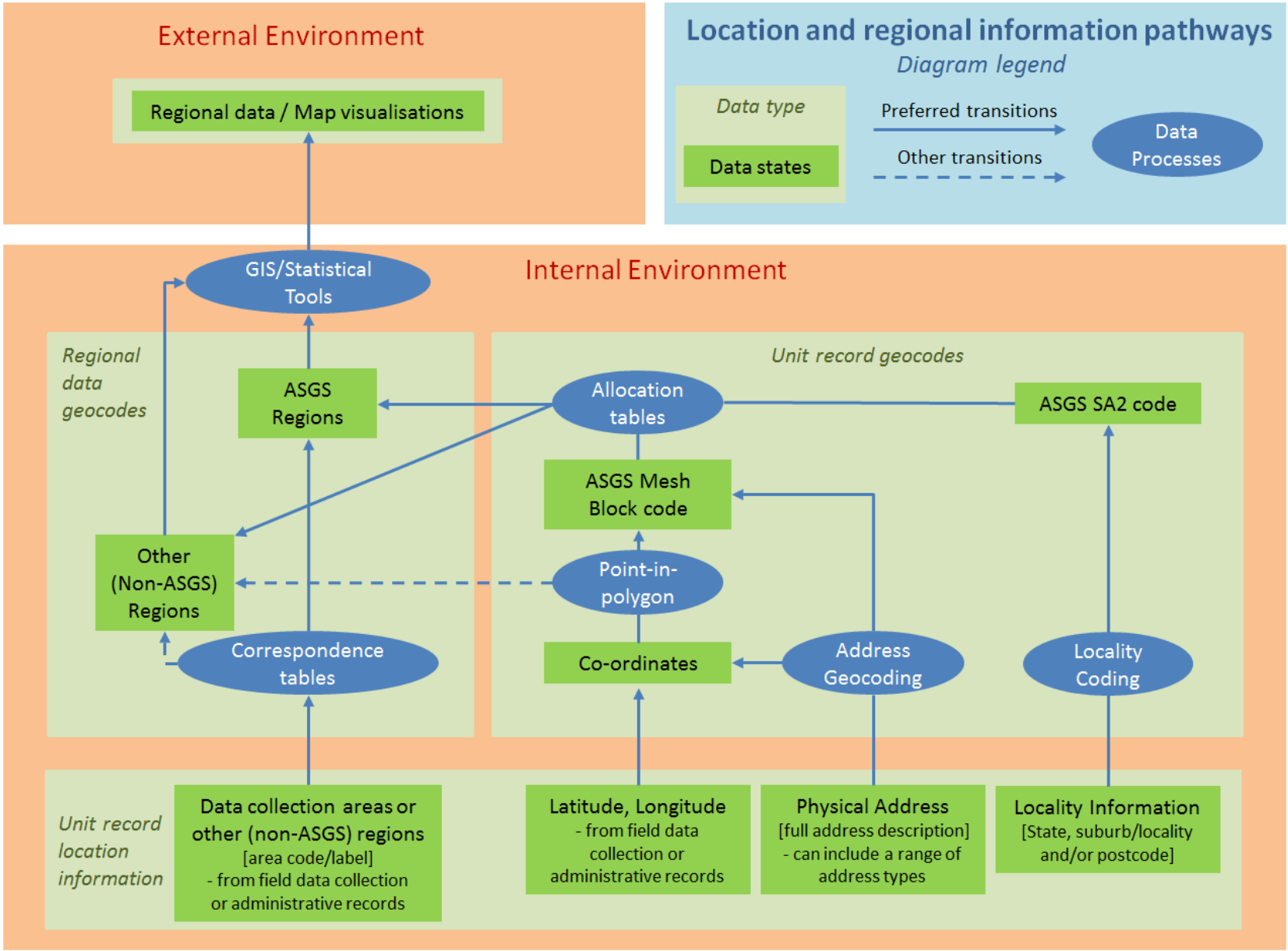
The geocoding method applied to a dataset will depend on the location information held in the unit record. In order of complexity for geocoding, the following location information can be used with the listed geocoding method:

- Location coordinates – coordinate and point-in-polygon geocoding.
- Full physical addresses– address geocoding.
- Partial physical addresses (i.e. suburb, postcode, state) – locality geocoding.

The following diagram describes the pathways for geocoding unit records based on these three options. The diagram also shows the use of data collection and region codes; these are the least preferred geocoding method for unit record data. They are discussed further in the SSF Guidance Material "Using Geographic Boundaries and Classifications with Statistics" on the NSS website.

[1] For more information follow this link: Statistical Spatial Framework (SSF)
[2] For more information follow this link: Australian Statistical Geography Standard (ASGS)

Diagram A – Location and regional information pathways



Geocodes in the Statistical Spatial Framework

Geocode – is a single location coordinate or a unique code that can be used to determine the position of a location on the Earth's surface. The unique code provides a direct link to a set of coordinates that defines a geographic object that represents that location – commonly a point or a polygon. The coordinates used must be related to a defined geospatial referencing system, such as the Geocentric Datum of Australia 1994 (GDA 1994).

For example, the location address "ABS House, 45 Benjamin Way, Belconnen ACT 2617" can have the following geocodes:

1. The location coordinate defined by the latitude: -35.2406 and longitude: 149.0678 (GDA1994).
2. The Mesh Block code - 80002993000; the Mesh Block that includes ABS House. This code directly references the polygon coordinate geometry that is associated with that Mesh Block, as defined by the Australian Statistical Geography Standard (ASGS).

Geocoding – is the process of assigning a geocode to piece of information (e.g. a unit record) using known location information, such as: a coordinate, address or locality/suburb. Geocoding processes are described in more detail in this paper.

For socio-economic datasets, geocoding usually involves assigning a geocode based the physical address for each statistical unit (e.g. persons, households, or businesses) in the dataset. If a detailed address is not available, the locality or suburb is often used to obtain a more general geocode. The SSF recommends that geocoding of socio-economic datasets be underpinned by the standards in the National Address Management Framework (NAMF)[3]. In particular, the SSF recommends use of the Geocoded National Address File (G-NAF)[4], which is the authoritative list of Australian addresses and locations coordinates. Use of NAMF and G-NAF ensures nationally consistent, standardised geocoding of address information.

Location coordinate – is a standardised latitude and longitude [5] for a physical address. This coordinate provides a high degree of precision, as well as providing flexibility to produce information for a range of current and future region types, as well as enabling other geographic uses in the future.

ASGS Mesh Block – is the smallest unit in the ASGS and is the building block for all the other ASGS units. Including a Mesh Block code as a geocode on a unit record enables data in the dataset to be released for all of the ASGS regions and other regions built up from Mesh Blocks. This can be done using a look up allocation table. ASGS regions are the common geography in the SSF. The ASGS reference in the data provides a location-based link between the data in the dataset and all of the other ABS data that are available for the ASGS regions, as well as data from many other datasets. This allows data from these sources to be directly analysed and compared with statistical data obtained from the dataset.

[3] For more information follow this link: National Address Management Framework (NAMF)
[4] For more information follow this link: Geocoded National Address File (G-NAF)
[5] Technical details: GDA 94 datum and unprojected geographic coordinate system - latitude & longitude.

Coordinate Geocoding

Datasets may already have location coordinates (i.e. latitude and longitude) attached to the unit records. These coordinates can be transformed into ASGS or other region types. This transformation uses Geographic Information System (GIS) point-in-polygon processes that allocate records based on their geocode to a region. Ideally, the point-in-polygon allocation process would also assign an ASGS Mesh Block code to each unit record. These Mesh Block codes can then be used to build-up data for all other ASGS geography using allocation tables, which map Mesh Block codes to larger geographic units.

Generally, the point-in-polygon allocation process is very accurate; however, the accuracy of the allocations is dependent on the following three factors:

- the quality of the original location information (i.e. address information);
- the geocoding process used to obtain the location coordinates for each unit record; and
- the precision or accuracy of the location coordinates included in the geocoding database.

Therefore, it is important to determine if the original location information (e.g. address information) reflects the location required for analysis or reporting (e.g. location of retail outlet versus head office, place of usual residence versus current residence). It is also important to understand which processes were used to obtain the geocodes and the accuracy of the coordinate information assigned. Documentation on the original location information and the geocoding process used should be obtained to make this assessment and, where necessary, expert advice sought.

Address Geocoding

If the unit record data contains full address information then address geocoding can be used. This method is much more accurate than locality geocoding, but the quality of the address information must be of a high standard for accurate and efficient geocoding.

Address geocoding is the process of matching a reported address against an index of geocoded addresses to obtain the geocode for the reported address. The Statistical Spatial Framework (SSF) recommends that the geocode information obtained from this process include a location coordinate, preferably obtained from G-NAF, and an ASGS region code, preferably an ASGS Mesh Block code.

There are a number of steps and processes associated with address geocoding, these include:

- Address collection
- Address repair
- Automatic address geocoding
- Assessment of the results
- Manual address geocoding and imputation
- Address Management

The ABS does not provide an address geocoding service, as there are a number of commercial organisations that offer geocoding services that incorporate G-NAF. More details on these companies can be found on the PSMA website.

Address Collection

The first step in address geocoding is obtaining address information for the statistical units (i.e. persons, households or businesses). Full street address is required for accurate and efficient geocoding that produces a location coordinate (i.e. latitude and longitude). Point-of-entry address validation processes can greatly improve the quality of address information obtained and the resulting geocoding of the information.

Address Information

An address is a textual description of a physical point on the surface of the earth. In Australia, it is usually either:

- a postal address to which mail is delivered, or
- a physical address, which is the actual location of a person, household or business.

A postal address location can be very different to a physical address location. For example, a rural property may have a post office box in the nearest town that is many kilometres away. In these and other similar instances, the postal address is not a good representation of where the actual property is located. Therefore, physical address is preferred to a postal address for identifying location.

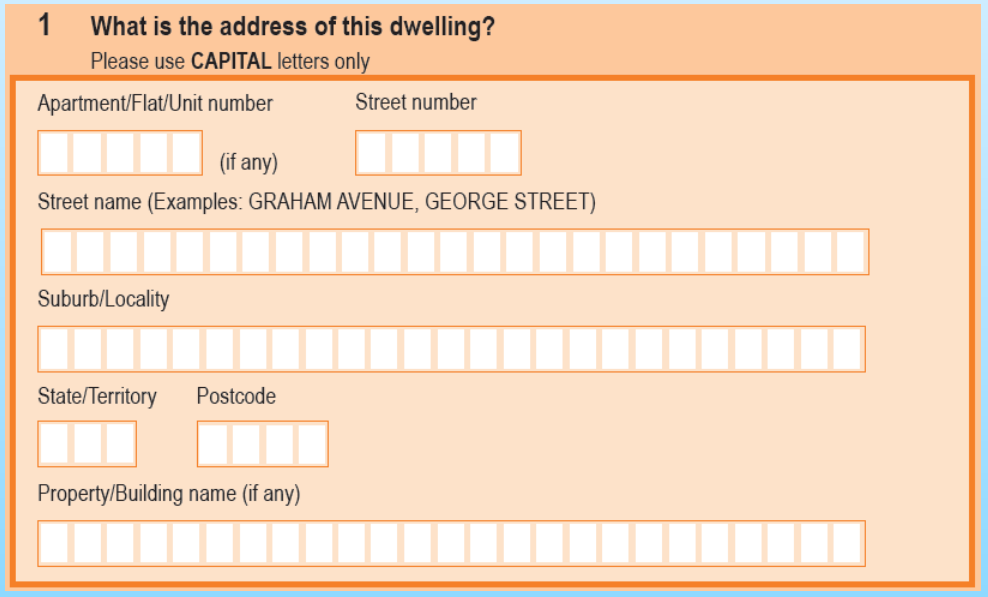
The following address elements are required to effectively capture the majority of physical addresses:

- Unit type and number
- Level type and number
- Building/property name
- Address number (can include textual prefixes and suffixes)
- Street/road name (or water feature or island name)
- Locality/suburb name
- State/territory name
- Australia Post Postcode

Each of these elements contains information that can improve the quality and efficiency of the address geocoding processes. Therefore, in any dataset that will be geocoded, it is recommended that each of these elements be collected and stored for each unit record. However, unit and level type is less important for this process and in some cases only the unit and level number is collected.

Collecting addresses information in the 2011 Census

For the 2011 Census of Population and Housing, the ABS collected address information on the paper form using the following response format.



A single physical location can be represented by a number of different address descriptions. For example, a rural address may be reported as a property name, a road name and number, or a lot number; but these address descriptions may all relate to the same dwelling or property. Alternatively, an urban dwelling may have several addresses used by different organisations to describe it or address mail to it. Some have only slight variations in the address but others may be very different, such as a corner block that has addresses for either of the streets where it is located. By collecting address information in a standard format (as outlined above) these types of problems can be minimised but the address geocoding system must also be able to interpret the complex information often used in addressing.

The custodians of official physical address information are the state & territory governments, and local government allocates official physical addresses. However, the official address may not be in common usage or be the only address in common usage for that property. The national authoritative list of physical addresses is the Geocoded National Address File (G-NAF); this dataset contains a range of addresses in use for properties across Australia.

Further information on address standards can be sourced from the following Australian Standards:

- AS/NZS 4819: 2011 Rural and urban addressing
- AS 4590: 2006 Interchange of client information

Address Validation

Depending on how addresses are collected, a point-of-entry address validation process can significantly improve the quality of reported addresses. This will then increases the quality and efficiency of the address geocoding process and reduce the time spent clerically correcting addresses after data entry.

At its simplest, point of entry address validation checks the reported address against an address index (or database) and, if there is a difference, asks the respondent to pick from one of the addresses in the index. By validating addresses at the point they are collected, against a national address list such as G-NAF, many simple address reporting errors can be avoided, such as spelling mistakes or incorrect locality names.

If the addresses in a dataset come from administrative sources, then it is important to obtain a good understanding of how the addresses were collected and the quality of the addresses information supplied. In this assessment, it is important to determine if the addresses contained in the dataset are physical addresses and reflect the location required for analysis or reporting (e.g. location of retail outlet versus main office, place of usual residence versus current residence).

Privacy

Address information on its own is not private information, as it does not directly convey any specific information about an individual or organisation; it is considered to be public domain information. However, privacy issues do need to be considered when address information is linked to name or other data (e.g. age, sex and income, or business type and turnover). Including the address can make the data identifiable to an individual person or organisation. When this occurs the information must be managed confidentially to maintain the privacy of the information. The SSF Guidance Material "Protecting Privacy for Geospatially Enabled Statistics: Geographic Differencing" on the NSS website provides guidance on privacy risks for custodians of geospatially enabled socio-economic data.

Address repair

Prior to address coding it may be advisable to examine the address records and repair or 'groom' the content of these records to facilitate the automatic address coding processes. The changes made will depend on the quality of the address information obtained, and will be informed by previous address coding processes undertaken with the system being used and the parsing capabilities of geocoding software. The impact of any changes made directly to the address information should be evaluated against the unrepaired addresses to ensure the changes do not inadvertently result in incorrect address matching.

Some address repair tasks that may be considered are:

- Correctly formatting fields for the address coder.
- Inclusion of state or territory on all records, based on postcode.
- Removing or examining incorrect characters in alphabetic and numeric fields.

Automatic address geocoding

Once addresses have been collected and repaired, the next step is the automatic address geocoding process. Automatic address geocoding is done by software that processes addresses in bulk and, where possible, allocates each address a coordinate (i.e. latitude and longitude) and, preferably, an ASGS Mesh Block code. Some geocoding systems will attempt to geocode the address to a larger ASGS region if a coordinate/ Mesh Block match cannot be obtained. The Statistical Spatial Framework (SSF) recommends that the geocoding system should use a G-NAF based coding index (or database).

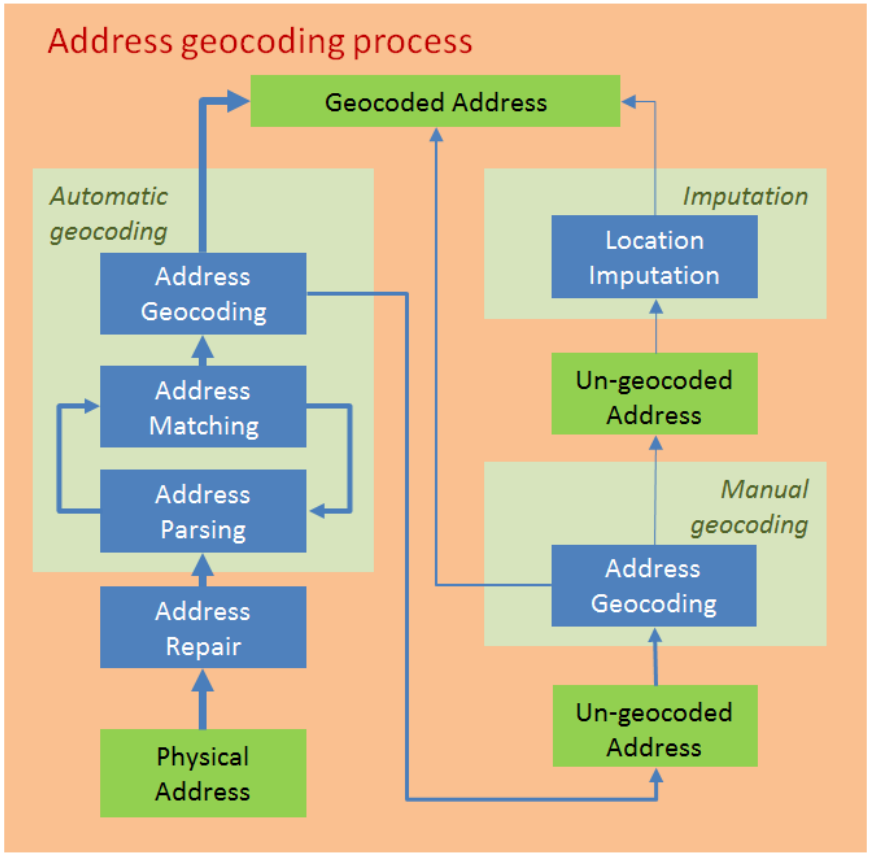
The automatic geocoding process is split into three separate components:

- Parsing – extracts relevant parts of the reported address into separate fields to allow matching.
- Matching – matches the parsed address information to a coding index.
- Coding – allocates a geocode and match information to each reported address.

This is an iterative process; the addresses may pass through these stages a number of times, until the highest quality match and geocode is found. With each pass the address information is parsed differently based on the range of possible matches identified from previous passes.

The diagram below shows the main components and information flows in the address geocoding process.

Diagram B - Address Coding Processes



Addresses geocoding results in the 2011 Census

For the 2011 Census of Population and Housing, the ABS allocated every address to a Mesh Block using address geocoding processes. Approximately 90% of all addresses were automatically geocoded using a G-NAF based coding index and the remaining 10% were geocoded to Mesh Block using manual coding techniques and imputation.

Generally, rural and remote addresses are more difficult to geocode than urban addresses. This is due to a number of factors such as:

- Poor reporting of addresses – a respondent may report a property name or a postal address.
- Geocodes not as accurate – may only provide a street or road mid-point coordinate and not a specific address coordinate.
- Rural addresses are not consistent across Australia – some rural addresses still use lot and section numbers (this is in the process of being standardised).
- Missing addresses – the address index may not be complete in some rural areas, or include the full range of addressing options for each physical location.
- Remote area addresses often have not been gazetted by State and Territory authorities (e.g. indigenous community addresses).

It is also worth noting that for any address index there will always be a gap between when new housing or development is completed and occupied, and when these addresses appear in the index. Therefore, addresses for newly developed areas may also be difficult to geocode.

Assessment of the results

It is important that an address geocoding system provides sufficient information to evaluate the results output by the system. This should include documentation on processes and logic used to geocode the address information.

For address parsing and matching, the system should provide information on the processes applied to each address. This information should indicate whether a match has been made, the degree to which the reported address has been matched to an address in the index, and what (temporary) modifications were made to the reported address to enable a match.

For the address coding stage, the information should indicate the accuracy of the location coordinate obtained for each address. For example, in G-NAF, a coordinate for an address is very accurate if the address is associated with a cadastral parcel, which is the smallest possible area a geocode can be determined for; and it is much less accurate when only associated with a street mid-point, or a locality. Coordinates based on street mid-points and localities can place the address a considerable distance from the actual location, which may distort the data from a geospatial perspective by placing the data in the wrong region.

Careful consideration needs to be given to what level of accuracy is acceptable at each stage of the geocoding process, as this can have a significant effect on the overall quality of the geographic or region information that can be obtained from the dataset. Where information from the dataset is required for large regions then lower levels of accuracy may be acceptable. If information is

required for small regions then high levels of accuracy will be required, which is likely to involve greater amounts of clerical intervention to obtain full geocoding of the dataset.

Manual address geocoding and imputation

Addresses that do not automatically geocode at all or to an acceptable standard must be geocoded clerically. Clerical processes may involve the use of a number of different techniques, sometimes in combination, such as:

- Deriving geocodes based on nearby addresses – within the same street or in the order that they were collected in the field.
- Manually searching maps, satellite imagery and websites to obtain geocodes or more detailed address information.
- Imputing a location using available information and probability methods – an optional final stage, if all addresses must be geocoded.

Deriving and imputing geocodes in the 2011 Census

For the 2011 Census of Population and Housing, private dwellings that had an incomplete address or no address had a Mesh Block code derived from adjacent dwelling addresses listed in the collection records. If a Mesh Block code was unable to be derived from the dwelling address then it was imputed into a Mesh Block located within the relevant collector's workload area. Imputation of the Mesh Block code used a probability proportionate method, based on distributions of already geocoded dwellings across the Mesh Blocks within the collector workload.

The 2011 Census of Population and Housing Dictionary provides more information on the derivations and imputations methods used.

Information Management

Managing the information associated with geocoding requires some consideration and planning. Most database management systems will meet the basic requirements for data storage and access. A Geographic Information System (GIS) may be required if more complex geospatial processing needs to be utilised, during and after data processing.

When storing the information obtained from geocoding addresses, apart from the coordinates and the ASGS region codes, consideration also needs to be given to what metadata from the geocoding process needs to be stored. The table below lists some of the basic information from the geocoding process that should be stored.

Address Information	Description
Location coordinate	3 data items should be stored: Latitude (x) – unprojected coordinates in decimal degrees using the Geocentric Datum Australia 1994 and precision to 8 decimal places. Longitude (y) – unprojected coordinates in decimal degrees using the Geocentric Datum Australia 1994 and precision to 8 decimal places. Elevation (z) – height above mean sea level using Australian Height Datum. (Note: the elevation may not be currently available, however, it is recommended to make allowance for it as it is expected to be a future requirement.)
Region code	The region code from the geocoding process – e.g. ASGS Mesh Block code.
Region reference information	Region classification and edition associated with the region code – e.g. ASGS2011_MB.
Address match reference	A unique code from the coding database for the address record that the reported address was matched to – e.g. for geocodes obtained from G-NAF this is the Persistent Identifier (PID).
Geocode source	The source of the geocode for each address – e.g. G-NAF, hard copy map or internet mapping.
Geocode confidence	Indicator(s) from the geocoding software of the accuracy or confidence level associated with the geocode assigned to each address.
Geocode software	The name and version of the software used to geocode the address.
Geocode index	The name and version of the coding index used.
Logs	Most geocoding software produce parameter and coding logs and it is best practice to retain them. This may need to be done separate to the unit record files.

Consideration should also be given to address information management; that is, how to store, process and maintain the address information in the address coding index. The maintenance of address information is an important ongoing issue for any geocoding system and the following issues should be considered:

- Methods for identifying new addresses not in the address index and methods for including these in the index for future use.
- Applying updates to the coding index from the supplier of the index.
- Managing improvements to geocoding software and supporting infrastructure.

Locality Geocoding

If only partial address information is available (such as suburb/locality, state and/or postcode), this may be used to geocode unit records to ASGS SA2 units or higher-level ASGS regions, and possibly other locality based regions. Suburb/locality, postcode and state are all basic parts of an address. This information can be used in conjunction with a suburb/locality to SA2 coding index to effectively geocode unit record data to the SA2 level and above.

> Locality coding indexes are available from ABS Geography Section by emailing: geography@abs.gov.au

For data that contains only postcode information, or another large-area region type, then accurate geocoding is generally not possible and the use of correspondences is usually the best option to convert data to other region types. It should be noted that, using correspondences will always result in less accurate data for the region being converted to, compared with the more direct methods discussed elsewhere in this paper. This loss of accuracy is due to the differences in coverage of any two different region types and the need to estimate a proportional redistribution of the data to the new region, based on the physical area of the new regions or the distribution of the population within the new regions.

> For more information about correspondences, see the SSF Guidance Material "Using Geographic Boundaries and Classifications with Statistics" on the NSS website.

Glossary

A glossary of geographic and related terms used in this paper is available on the ABS website.

Where can I get further information?

More information on the Statistical Spatial Framework can be found on the SSF webpage on the NSS website.

Information about privacy and geospatial information is contained in the SSF guidance material paper, "Protecting Privacy for Geospatially Enabled Statistics: Geographic Differencing" on the NSS website.

Information about using geographic regions is contained in the SSF Guidance Material "Using Geographic Boundaries and Classifications with Statistics" on the NSS website.

Any questions or comments on this paper or other statistical geography topics can be emailed to geography@abs.gov.au