**UNITED NATIONS SECRETARIAT**
**Department of Economic and Social Affairs**
**Statistics Division**

**ESA/STAT/AC.115/5**
**May 2007**
**English only**

---

**United Nations Expert Group Meeting on**
**Contemporary Practices in Census Mapping and**
**Use of Geographical Information Systems**
**29 May-1 June 2007**
**United Nations, New York**

# How to Geocode Information in Population and Housing Censuses*

Prepared by

Demographic and Social Statistics Branch
UN Statistics Division

* This document is being reproduced without formal editing.

## How to Geocode information in Population and Housing Censuses[1]

### *Summary and introduction*

1. This paper is an abridged introduction[2] to the broad practice of geocoding. It serves primarily to introduce some concepts and to foreshadow the preparation of 'subsidiary papers" on two alternate methods of acquiring geocodes for units in a population census.
2. As such this paper is targeted at more senior managers in statistical agencies. The subsidiary papers will contain more technical details and thus be aimed at more specialised staff undertaking the development and/or operational tasks associated with implementing geocoding.
3. A question that may underlie reading this document is when a National Statistical Office should consider transition from traditional, paper based, mapping systems to the use of geocoding as part of a digital basis for mapping activities required by the census and other statistical activities. The answer to this question can only be considered country by country. In part it will be informed by the availability of resources: staff with the necessary skills; the range of hardware and software required; and funds to undertake the required investment in capturing the base data. It should also be driven by user demands for statistical information that can be provided more effectively and efficiently using geocoded data.

### *Background and context*

4. The actual task of geocoding is a technical activity. However it is important to place this in the context of the broad setting of a National Statistical System. The following sections address this from the perspectives of:
    - the underlying net benefits of the approach;
    - the classifications required to support a geocoding operation; and
    - the institutional placement of a Census operation within an institutional framework.

---

[1] The initial drafting of this paper was undertaken by staff of the United Nations Statistical Division. Helpful comments on that draft have been received from Angela Me (UNECE), Alister Nairn (Australian Bureau of Statistics), Bolaji Taiwo (UNFPA, Kabul), Shawn Hanks and Charlene Leggieri (US Bureau of the Census), Roberto Bianchini (Universitá di Roma) and Liz Gavin (Statistics South Africa).

[2] For a more detailed coverage of geocoding, especially in relation to a census of population and/or housing see United Nations , 2000, "Handbook on geographic information systems and digital mapping" Studies in Methods series f No. 79; p187.

## Cost-Benefit analysis.

5. There are always a range of benefits, especially longer term, from adoption of geocoding. However it is also the case that the initial work required to set up the system will involve additional short term costs. It is important that an agency planning to undertake a first geocoding exercise undertakes a comprehensive assessment of these costs and benefits. This may well show that a staged approach, through which some of the activities are deferred, is preferable to a full scale "everything-at-once approach".
6. It is not the aim of this document to provide the details of the analysis to be performed. However it will be assumed that an agency has examined the impact of the geocoding exercise on agency budgets and has a well developed understanding of the user needs being met by the additional work.

## Classifications required

7. Within a statistical system the key application of the codes which result from a geocoding process is to place the coded units within one or more classifications of geographic entities. Typically a country will be divided into a number of spatial entities conveniently referred to as civil divisions. There will, in most cases, be more than one level of civil division giving schemes such as those illustrated below.

### Table 1: Examples of hierarchies of Civil Divisions

| Examples of possible arrangements | | | | |
|---|---|---|---|---|
| | **a** | **b** | **c** | **d**[3] |
| First | Region | Region | Province | State |
| Second | Department | District | Municipality | Statistical Division |
| Third | Arrondissement | Town/village | | Statistical sub-division |
| Fourth | Canton | | | Municipality |

8. There will also be other classifications applicable to the geographic disaggregation of a country[4]. Some of these will be compatible with one or more of the units

[3] In some countries the national statistical agency may form sets of units, such as these, from the other examples shown as standard elements of statistical infrastructure. In time, these sets may themselves become standard units of civil administration.

[4] For an example of a more detailed description of the units that are considered in one country's statistical geography see Australian Bureau of Statistics "Australian Standard Geographical Classification" Bulletin

shown above (for example a Health Service administered by a State Government will use administrative boundaries that fit within State borders, but not necessarily respect the boundaries of Municipalities within that State). In other cases the alternate geographies may only coincide with the National boundary and cross even first level Civil Division boundaries.

9. In each case the boundaries of the areas can be defined as a set of latitudes and longitudes within which the geocoded units will be placed. To a large extent forming these classifications is an essential facet of the geocoding exercise. It represents the way in which the actual latitudes and longitudes are used. It is suggested that creating a set of digitized boundaries for the key Civil Divisions is a vital first step in a geocoding exercise. As coding of individual responses to a pre-defined classification is necessary for producing aggregated statistics, so the geo-coding of the location at which information is collected provides for aggregate statistics associated with a particular geographical unit.

10. Increasingly, users of statistics are interested in using statistics at the lower levels of this set of geographic classifications. In the context of census information, a key attribute of this mechanism for data collection is the ability to provide results with low/zero levels of sampling error for small areas. In view of these needs of users of statistics it is crucial that the geocoding process places the coded units correctly in space relative to these civil divisions.

11. There may also be more detailed levels of statistical geography (for example individual enumeration areas or block-level units) than those shown in Table 1. The general principles of coding dwelling units to fit within the boundaries of a geographic area can readily be extended to these smaller disaggregated areas. It is emphasised this is a crucial underlying attribute of the design of these very detailed units. Since the nature of the very detailed units is likely to vary between countries and to be established by specific user or operational requirements and the availability of technology to code to these levels they will not be considered further in this paper.

### *Institutional framework*

12. In the same way as the previous sub-section placed the geocoding exercise in a statistical context it is important to recognize the institutional framework within which the geocoding is undertaken.

---

1216.0 (or http://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/1216.0Contents1Jul%202006?opendocument&tabname=Summary&prodno=1216.0&issue=Jul%202006&num=&view= ). The range of units classified are shown in the structural chart in Chapter 1. Note that the lowest unit "Census Collection District" is equivalent to Enumeration Area.

13. In most cases [5] National Statistical Offices do not have a core objective of measuring and collecting spatial data or developing specialized mapping tools. This is more usually the function of a specialist organization known, in the past by a name such as "National Mapping Office" but now becoming known by names such as "National Land Information Office". Equally the mapping offices do not have expertise in the statistical aspects of geocoding.

14. Clearly what is needed is a cooperative arrangement between the two agencies and/or other agencies which maintain spatial data and have expertise in this domain[6]. The details of arrangements to be implemented are beyond the scope of this paper, but should include an agreed ownership or custodianship of the geocodes and the geocoded data and responsibility for maintaining and updating these data.

15. Depending upon the practices followed in each country it may be desirable to formalize this cooperation through a contractual arrangement (under which the NSO contributes to the cost of work by the other agency) or to simply operate under a co-operative, or administrative, agreement between the agencies perhaps supported by the legislation establishing the agencies). In either case the other agency/agencies should be made aware of the potential benefits to them of participating in the scheme. These benefits could include:
    - the existence of an agreed consistent and standardised set of geocodes for a country will maximise consistency of decision making of matters based on spatial analysis;
    - obtaining statistical information about the areas covered by the maps;
    - as well as receiving feedback on conditions "on the ground" from a huge field validation exercise (under most operational models of census taking).

## *Definitions*

### Conceptual Definition of Geocoding

16. The conceptual; definition of geocoding should cover 2 situations:
    a. a Geographic Information System function that determines a point location based on an address. It could generally be expected that such point locations will be relatively precise (eg $\pm$ 2 metres) in accuracy and will often be based upon use of Global Positioning System technology.
    b. the more general process of assigning geographic codes to features in a digital database.

[5] Mexico and Brazil are examples of countries in which the cartographic and statistical functions are combined in a single National agency.

[6] By way of example in a number of African countries agencies with responsibilities for environmental issues have taken the lead in digital collection of spatial information.

### *Operational definition*

17. Geocoding is the computer oriented process which converts information about a unit from which statistical information is collected into a set of coordinates describing the geographic position of that unit.

18. Under the first element of the conceptual definition this may involve collecting precise data at the level of point locations (or very low geographic level such as a city block) to assign geographic codes for use in the dissemination of output from the statistical exercise.

19. In a Population and Housing Census the unit actually coded may be a block of land, a building or (preferably) a dwelling unit. The coordinates may be taken as the center of the unit (often the case when the coding is based upon cadastral information) or some other convenient point such as a specified corner of a building or the point of entry to the building.

20. The crucial element is that the coordinates contain the latitude and longitude of the unit, enabling its position relative to the surface of the earth to be expressed in a standard way. In some implementations of geocoding a third coordinate – effectively, altitude - may also be used, for example where a multi-storey building comprises several dwelling units which can all be allocated the same latitude and longitude. User demands for such additional information will usually be satisfied by identifying the "floor" of the building on which the dwelling unit is located. As there is no standard for such a classification[7] this will need to be resolved by each country as required.

21. With regard to the second element of the conceptual definition, it is necessary to develop a full set of codes covering each geographic unit to be classified. Where the most detailed units (eg Enumeration Areas, Mesh Blocks) are combined in various ways to form less detailed units concordances should be used to relate the various combinations.

### *Relationship to other census processes and other elements of the National Statistical system*

22. At a broad level of detail geocoding is central to the entire census operation. All operational processes are undertaken in relation to the units within a spatial entity often referred to as an Enumeration Area (EA).

23. Meeting the objectives of the census in terms of satisfying users' needs for information is also performed according to the users' specification of a spatial unit of interest. This may range from the country as a whole down to the Enumeration Area (or, in some countries smaller areas referred to as mesh blocks or unit blocks). In some discussions reference is made to the geocoding of information for the

[7] For example, British usage has the 1st floor as the one above ground level whereas in the US the first floor is at ground level. In some cases, where land is steeply sloping it may be difficult to even identify ground level.

individual dwelling units as meeting user requirements but in most countries it is not possible to disclose information at that level due to the need to preserve confidentiality.

24. Although beyond the strict ambit of this topic, it could be noted geocoding can also be of benefit to other elements of a National Statistical System. Benefits could include:
    a. the improved maps often generated through geocoding will benefit tasks relying on field work such as household surveys.
    b. It may be possible to use the point locations to produce sub-national output for customized geographies (where the sample size and required geography are compatible is sufficient).
    c. the formalization of coding schemes will be of use to all statistical endeavours using the coding schemes.

## Approaches to data collection

25. At a very broad level two alternatives may be considered.

26. The first approach can be termed the 'direct collection approach' since the coordinates are directly collected for each building in the country, either by digitizing from available small-scale topographic and city maps or by collecting the coordinates using field techniques. If it is necessary to obtain the third attribute of elevation each dwelling unit would have to be visited and the elevation recorded.

27. The second approach, termed the "matching approach" can be employed where a comprehensive street network database and a master file of addresses of the population exist. With this information, GIS can locate any given address on the street network in a process known as *address matching*. This approach does not require the precise coordinates of point location and can be considered an example of the general geographic coding covered by the second element of the conceptual definition of geocoding.

28. A variation on the matching approach could be employed where it is possible to identify the cadastral reference of a plot of land and by matching these details to a cadastral register containing the spatial coordinates obtain the required latitude and longitude.

29. With either of these matching approaches further complexity would be required to obtain the elevation for each dwelling unit and it may be impossible where the target database does not separately identify the levels with in a building..

30. In some circumstances it may be desirable to merge the two approaches. For example initial geocoding of the location of dwellings may be undertaken by direct collection but text strings containing addresses from census questions such as place of work or place of residence 'n' years ago may be matched against that initial list.

## Preceding tasks and conditions

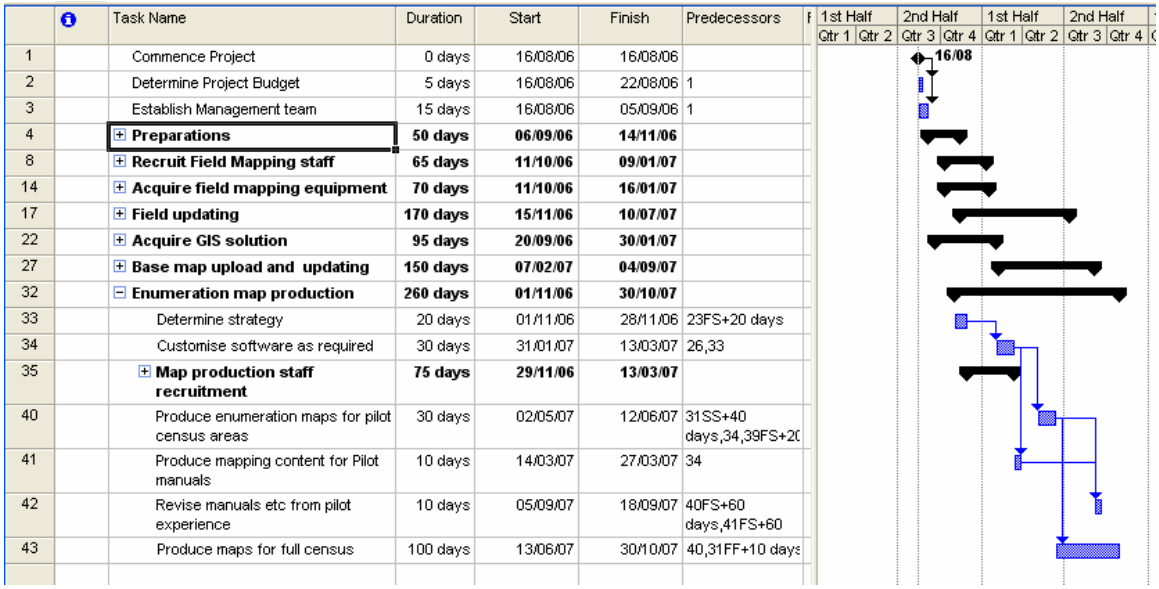
31. The specific tasks and conditions needed as pre-requisites for the two approaches are rather different. They will be described in detail in forthcoming papers "9.1.1 Geocoding by Direct Collection" and "9.1.2 Geocoding by Matching".

32. It should be noted that for many of the benefits of either approach to be gained there needs to be a digitally registered base map for the country against which the geocoded units can be presented or analysed. This could be a raster backdrop or a fully vectorised digital spatial dataset, depending upon the uses intended.

33. The following diagram shows the highest level activities in a Gantt Chart drawn up to reflect a possible approach to preparing the hard copy mapping materials for a census enumeration using a digital base map. Annex 1 is a copy of the complete Chart. If the enumeration is to be undertaken using Personal Data Assistants or some other form of handheld device with built in mapping capability the tasks relating to paper map production in the Chart should be replaced with tasks relating to downloading the mapping information from the base map to the devices. It is likely that the time required for system development and testing of the electronic approach will be longer than the hard copy approach, but the time for the reproduction of the material will be reduced.

34. It is stressed that the timelines in this Chart are notional, and the set of tasks is intended only to suggest to countries the types of tasks they should contemplate in designing a digital map based census. In particular countries may find it useful to include timelines for undertaking field work in defined parts of the country and thus monitor the progress of the operation at a finer level than is suggested by the rather blunt set of tasks shown here.

## Diagram 1: Gantt Chart showing activities in digital census mapping



| | Task Name | Duration | Start | Finish | Predecessors |
|---|---|---|---|---|---|
| 1 | Commence Project | 0 days | 16/08/06 | 16/08/06 | |
| 2 | Determine Project Budget | 5 days | 16/08/06 | 22/08/06 | 1 |
| 3 | Establish Management team | 15 days | 16/08/06 | 05/09/06 | 1 |
| 4 | **Preparations** | **50 days** | **06/09/06** | **14/11/06** | |
| 8 | **Recruit Field Mapping staff** | **65 days** | **11/10/06** | **09/01/07** | |
| 14 | **Acquire field mapping equipment** | **70 days** | **11/10/06** | **16/01/07** | |
| 17 | **Field updating** | **170 days** | **15/11/06** | **10/07/07** | |
| 22 | **Acquire GIS solution** | **95 days** | **20/09/06** | **30/01/07** | |
| 27 | **Base map upload and updating** | **150 days** | **07/02/07** | **04/09/07** | |
| 32 | **Enumeration map production** | **260 days** | **01/11/06** | **30/10/07** | |
| 33 | Determine strategy | 20 days | 01/11/06 | 28/11/06 | 23FS+20 days |
| 34 | Customise software as required | 30 days | 31/01/07 | 13/03/07 | 26,33 |
| 35 | **Map production staff recruitment** | **75 days** | **29/11/06** | **13/03/07** | |
| 40 | Produce enumeration maps for pilot census areas | 30 days | 02/05/07 | 12/06/07 | 31SS+40 days,34,39FS+20 |
| 41 | Produce mapping content for Pilot manuals | 10 days | 14/03/07 | 27/03/07 | 34 |
| 42 | Revise manuals etc from pilot experience | 10 days | 05/09/07 | 18/09/07 | 40FS+60 days,41FS+60 |
| 43 | Produce maps for full census | 100 days | 13/06/07 | 30/10/07 | 40,31FF+10 days |

35. It is intended that other Gantt Chats will be included in the subsidiary papers showing the direct and matching approaches.

36. The following table provides a conceptual starting point for contrasting the requirements of the two approaches against a range of prerequisites. It will be supplemented by similar material in more detailed papers.

### Table 2: Requirements for approaches to Geocoding

| Task/condition | Direct collection | Matching |
|---|---|---|
| Existence of digital base map for country | Highly desirable | Highly desirable |
| Statistical staff with expertise in use of GPS | Essential | Not important |
| Acquisition of large numbers of GPS receivers | Essential | Not important |
| Geo-referenced list of addresses or equivalent | Not important | Essential |
| Excellent address matching algorithms | Not important | Essential |
| Existence of a rational, consistent, and locally-recognized addressing system for housing units | Highly desirable | Essential |

## Dependent tasks and outputs

37. The broad outputs for the two approaches have some similarities, because the overall objective of the process should be to compile some precise spatial information about the basic units in the collection. However, at a finer level of detail there will be differences due in large part to differing user requirements..

38. The downstream dependencies and outputs will be described in detail in papers "9.1.1 Geocoding by Direct Collection" and "9.1.2 Geocoding by Matching".

## *SWOT analysis*

*39.* In the final paper there will be entries under each of the following contrasting (a) detailed geocoding against basic geocoding (ie just showing an EA code against the units) and (b) contrasting the two major alternatives. The following are just a smattering of the points that have first occurred to illustrate the approach.

## Strengths

40. The major strengths of geocoding units in a census of population and housing are that it will enable:
    a. early (or often the first) steps towards a fully GIS based approach to census mapping;
    b. presentation of high quality maps for use in the collection phase;

c. reduction of the work required in updating maps for future censuses; and
d. the aggregation of records into customized units for satisfying users' requirements.

41. The inclusion of housing unit structure locations can improve the quality and usability of maps used in the collection phase of a census. However the depiction of individual housing unit locations on maps may also clutter the map or otherwise obscure detail that is needed for census collection. The decision to include housing unit structure locations should be weighed against the needs of the field workforce and cartographic considerations.

42. The strength of the direct collection approach is that it does not require some other agency to have already developed a fully geo-referenced list against which units can be matched. In contrast, the strength of the address matching approach is that it will be far cheaper for the statistical agency to perform since the costly field work has already been done.

## Weaknesses

43. In both cases of geocoding there is a need for relatively highly trained staff. It is very common for inexperienced staff using GPS technology for the direct collection approach to make errors. Under the matching approach it is necessary to collect very detailed information about the addresses to avoid duplication: for example
    - in Manhattan the prefix E or W to Street addresses is crucial;
    - in New York there are many addresses such as 21st Street which could be duplicated without a zip (postal) code to differentiate between addresses in various (adjacent) Boroughs;
    - many urban areas include thoroughfares which can be differentiated only by the suffix (eg Road, Street, Avenue, Place etc)

44. A weakness of geocoding is that in most cases it places great reliance on an external provider. In this application the reliance is upon the providers of the base maps to have provided high quality material as input to the process.

45. The relative weakness of the 'traditional' approach to census cartography, relying heavily on field work and clerical updating of paper maps is that it requires a massive workload each census and is much harder to provide quality assurance since that can, almost by definition, only be done by further field work.

## Opportunities

46. The main opportunity flowing from geocoding is that it makes a major step forward in national infrastructure relevant to many fields of work beyond statistical applications. As well as the increase in well-being that would flow from this it also presents an opportunity for focusing cooperative activities between a number of agencies.

47. A further opportunity is the flexibility for statistical output added by the existence of geocoded records. Geocoded records can be classified to any user-defined scheme of geographical units. Particular note should be made of the ability of point

geocoding to enable recoding of data, without field work, when administrative boundaries change.

### Threats

48. Unless there is a high level support for the undertaking there is a risk of difficulty in communication between the stakeholders in a geocoding exercise leading to contention for ownership of the end result.

49. The major threat from a continuation of a traditional approach is that the amount of field work is so great that it may fall behind leading to the census itself having to be postponed.

## *Case Studies of Best Practices*

50. Case studies of best practices will be referenced in the papers describing the more detailed practices.

## *References*

Australian Bureau of Statistics, 2006, "Australian Standard Geographical Classification" Bulletin 1216.0

United Nations, 1998, "Principles and Recommendations for Population and Housing Censuses" Statistical Papers series M No. 67 Rev.1

United Nations, 2000, "Handbook on geographic information systems and digital mapping" Studies in Methods series F No. 79;

United Nations, 2001, "Handbook on Census Management for population and Housing Censuses" Studies in Methods series F No. 83 Rev.1