Statistical Commission
Fifty-second session
1–3 and 5 March 2021
Item 4(j) of the provisional agenda
Items for information: Integration of statistical and geospatial information

Background document
Available in English only

The Global Statistical Geospatial Framework: Implementing Geocoding

Prepared by the United Nations Expert Group on the Integration of Statistical and Geospatial Information

The Global Statistical Geospatial Framework: Implementing Geocoding

Introduction

This Scoping Paper has been developed by the United Nations Expert Group on the Integration of Statistical and Geospatial Information and is submitted to the Statistical Commission as a background document to its report [1] to 52nd session in March 2021. This Scoping Paper aims to provide straightforward, practical guidance on how countries can develop their ability to geocode statistical unit record data, highlighting resources that can be used. This document is part of a series of Scoping Papers currently being developed by the Expert Group, as part of their mandate to help countries adopt and implement the Global Statistical Geospatial Framework (GSGF).

Relevant Principles of the Global Statistical Geospatial Framework

Principle 1: Use of **fundamental geospatial infrastructure** and **geocoding**
Principle 2: **Geocoding unit record data** in a data management environment

What is Geocoding?

Generally, people prefer to use descriptions of locations instead of coordinates to navigate their environment. As an example, for the delivery of goods, we supply an address instead of the coordinates of our doorstep. However, modern geospatial technologies depend on absolute position data coordinates within a specific reference system, rather than relative location descriptions. The process of bridging this divide is referred to as geocoding.

Geocoding is the method of linking a description of a location to the location's measurable position in space. Geocoding links unreferenced location information (e.g., an address, or other location description) associated with a statistical unit (e.g., housing unit or business) to a set of coordinates within a coordinate system [2]. These resulting coordinates are the geocode. More formally stated, geocoding is generally defined as the process of geospatially enabling statistical unit records or other nonspatial data (such as address lists or housing unit records) by creating x- and y- (and potentially z) coordinates [3] and linking them to each record. Once geocoding is performed on individual statistical unit records, they (or the associated data) can be aggregated into larger geographic units (e.g., states, provinces, or municipalities) for statistical analysis. The records are ready for further applications such as methodologies to ensure confidentiality and avoid data disclosure.

Geospatial science is a rapidly growing field of study, with an evolving body of geospatial terms and concepts as the technologies are adopted in a wide variety of applications. Geocoding is often referred to with the terms geoenabling, geolocalising, or simply "linking" in some implementations. Frequently geocoding is associated with, and considered a subset of georeferencing. Georeferencing, in its broadest definition, is understood to be the process of linking geospatially enabled data to a common geospatial reference frame that allows geospatial presentation and analysis of those data, usually in Geographic Information System (GIS) software. Georeferencing requires linking coordinates to a defined geospatial

[1] E/CN.3/2021/27
[2] Also refered to as a spatial reference systems.
[3] x- and y- coordinates referring to a Latitude and Longitude or an Eastings and Northings, with the z- coordinate referring to elevation are the most commonly used, but other references are in use.

reference frame (i.e. a geospatial datum, ellipsoid, coordinate system, and often a projection). Georeferencing may refer to the alignment of orthoimagery or digital copies of paper maps with their inherent geographic coordinates (i.e., geocodes); or the transformation of geospatial data from a one defined geospatial reference frame to another.

Why is Geocoding needed?

Appropriately geocoding statistical unit records to a specific geospatial location fosters the greatest opportunity to reuse and aggregate statistical data. The GSGF states that *"all statistical unit records should include or be linked to a precise geographic reference (an x- and y- coordinate), and if not, the smallest geographic area possible".* This recommendation for using an x- and y- coordinate for geocoding was first issued by the Expert Group in 2018[4] and is reiterated by the Expert Group in 2021.

By geocoding each statistical unit record in a consistent, accurate, and precise manner, aggregation and disaggregation of their associated statistical data by geospatial location becomes possible. In this manner, the dissemination of statistical data using common geographic areas is enabled, promoting the reuse and comparability of data throughout time. Subsequent principles and key elements of the GSGF guide the dissemination of geospatially enabled statistical data in-line with prevailing privacy and confidentiality concerns, and national and international norms, standards, and policies regarding data disclosure.

How can records be geocoded?

Modern geocoding processes are largely automated, involving matching captured data with a reference database with some in-built spatial intelligence to improve the matching process. The efficiency of geocoding relies on having a comprehensive reference database of addresses and locations. This is a component of a mature national spatial infrastructure. Geocoding is also helped by having a standardised, structured description of a location. For example, a street address contains a number of specific elements with formatting requirements that are used in geocoding.

Geocodes can be generated directly (i.e., coordinates accepted as being specific for the statistical unit record) or indirectly when they use an internal point of a geographic area. Conceptually the most accurate geocodes are the x- and y- coordinates that were assigned to a statistical unit record at the time it was collected by using a Global Navigation Satellite System (GNSS), such as the Global Positioning System (GPS) or coordinates from the nationally agreed geodetic reference frame. Equally specific are geocodes assigned using specific standardised structure IDs or even within structure IDs (e.g., one apartment within an apartment building). The next most specific geocodes are for addresses or standardised parcel IDs.

If this level of precision is not possible, geocodes can be generated using an internal point (e.g., a centroid) for any functional area which represents a specific geography (e.g. an enumeration unit, a small building block geographic area, or a small area grid cell up to localities, postal code areas, or even second-level administrative units). Regardless of the geographic level used to geocode a statistical unit record, the manner of geocoding must be consistently documented for each statistical unit record in a dataset along with a corresponding record of a time and date for each record when each record was geocoded.

[4] E/CN.3/2018/33 https://unstats.un.org/unsd/statcom/49th-session/documents/2018-33-GeoInfo-E.pdf

To identify available data to geocode statistical unit records or help identify their absence (and in turn, identify gaps where capacity can be developed), the 14 Global Fundamental Geospatial Data Themes[5] may be useful. These are 14 Themes considered fundamental to strengthening a national geospatial information infrastructure. Specifically, the Global Geodetic Reference Frame (x- and y- coordinates), Addresses, and Functional Areas are directly relevant to geocode statistical unit records.

x- and y- coordinates

Addresses

Functional Areas

The Expert Group urges geocoding in the most accurate way possible to allow the most flexibility in combining various geocoded data and reiterates its previous recommendation to geocode statistical unit records with specific x- and y- coordinates. If this is not possible, it recommends geocoding (creating x- and y- coordinates) by using Addresses, or lastly by using the smallest Functional Areas possible.

Further Reading and Associated Resources

- The Global Statistical Geospatial Framework E/CN.3/2020/25 http://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf
- Integrated Geospatial Information Framework Strategic Pathway 4: Data https://igif.un.org/
- Australian Bureau of Statistics: Geocoding Unit Record Data Using Address and Location: https://www.abs.gov.au/websitedbs/d3310114.nsf/home/Statistical+Spatial+Framework+Guidance+Material/$File/Geocoding+Unit+Record+Data.pdf
- Academic resources: (such as Texas A&M's geocoding resources https://geoservices.tamu.edu/Services/Geocode/
- The Global Fundamental Geospatial Data Themes: http://ggim.un.org/documents/Fundamental%20Data%20Publication.pdf

[5] E/C.20/2020/14/Add.1 http://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/E_C.20_2020_14_GFGDT.pdf and http://ggim.un.org/documents/Fundamental%20Data%20Publication.pdf