

NAACCR Geocoder Data Dictionary
(Texas A&M, NAACCR, NCI Geocoding System Automated
Geospatial Geocoding Interface Environment (AGGIE))
Last Revised: July 2017

Introduction

This document defines the data items generated by the NAACCR geocoder during the batch database geocoding process.

The user submits a database table as data input with a minimum of the following data items:

- Unique Record ID
- Street Address
- City
- State
- Zip Code

The database table can include more than these four data items, but only these four are required for the geocoding process. For example, the user may want to include Patient ID and Central Sequence Number as two additional data items for each record to uniquely link the record back to the original cancer database. The user can then easily create a unique 10-character record ID from the 8-character Patient ID joined with the 2-character Central Sequence Number. Or, the ID can just be a sequential number.

Including the diagnosis year in the database table can be useful so that the user can more easily divide the data into different subsets for linking with 1990, 2000, and 2010 census geometries.

During Step #3 of the batch database geocoding process, the user sees the six categories of data items that will be generated from the geocoder. The first two categories are always returned:

- **BookKeeping:** Administrative data items.
- **OutputGeocode:** Data items related to latitude and longitude.

The remaining four groups of data items are optional and will only be included if selected by the user:

- **CensusValues:** Data items related to census tract, census block, and census block group.

NOTE: In order to get the geocoded county, tract, block group, and certainty codes, a user must check this box.

- **ParsedAddress:** How the geocoder divided up the input address into its individual elements and standardized abbreviations.
- **MatchedAddress:** How the geocoder used the parsed address to make a match.
- **FeatureAddress:** The parsed address of the geographic feature to which the address was matched.

Data Item List

This section lists the 129 data items that appear as output in the geocoded database table. The data items are grouped in their six categories and are presented in the order that they appear in the table.

1) Book Keeping Data Items

Source: This is the source of the geocoded data. The default is “USCGeocoder” as the software was originally developed at the University of Southern California.

UpdatedGeocoding: Indicates the status of the geocoding. 1= Completed
Blank = Not completed

TimeTaken: The time it took the geocoder to process the record in units of 100th of a second.

TransactionId: A unique alpha-numeric ID assigned to the record.

Version: The version number of the geocoder software.

NOTE: As of July 18, 2017, the current version is v4.03.

ErrorMessage: Any errors encountered during the geocoding process are indicated here.

2) Output Geocode Data Items

Latitude: The latitude of the address in units of decimal degrees. The datum of the decimal degree is North American Datum of 1983 (NAD 83). Latitude north of the equator is positive. Note that even though the output has 13-digits in the decimal place, the precision of the location is far less precise.

Longitude: The longitude of the address in units of decimal degrees. The datum of the decimal degree is North American Datum of 1983 (NAD 83). Longitude west of 0 degrees (the Prime Meridian) and east of 180 degrees (approximately the International Date Line) is negative – this applies to the entire North American continent with the exception of the tip of the Aleutian Islands in Alaska. Note that even though the output has 12 digits in the decimal place, the precision of the location is far less precise.

MatchType: The method used to match the address:

- Exact
- Nearby:Relaxed
- Relaxed
- Relaxed;Soundex

MatchedLocationType: The type of address being geocoded to:

- LOCATION_TYPE_POST_OFFICE_BOX
LOCATION_TYPE_RURAL_ROUTE
LOCATION_TYPE_STREET_ADDRESS
- LOCATION_TYPE_USPS_ZIP_CODE (for ungeocodable street addresses)

FeatureMatchingResultType: Indicates if the feature matched to was exact or nearby:

- Nearby
- Success

InterpolationType: Indicates the type of interpolation that was used in determining the geocode:

- ArealInterpolation (for building centroid, parcel, city, or ZCTA [ZIP Code Tabulation Area])
- LinearInterpolation (for street segment)

FeatureMatchingGeographyType: Indicates the type of feature used in determining the geocode:

- BuildingCentroid
- City
- Parcel
- State
- StreetSegment
- ZCTA

FeatureMatchingHierarchy: Indicates the hierarchy used in determining the geocode. When the hierarchy is FeatureClass, the geocoder will choose the first geocode that matches in the following hierarchy order: Exact parcel centroid, Nearest parcel centroid, Uniform lot interpolation, Address range interpolation, ZIP code centroid, City centroid, County subdivision centroid, County centroid, Country centroid:

- FeatureClassBased
- UncertaintySingleFeatureArea
- UncertaintyMultiFeatureGraviational
- UncertaintyMultiFeatureTopological

The uncertainty hierarchy directs the geocoder to choose the geocode with the lowest uncertainty as the resulting “best geocode” that should be returned for an address.

FeatureMatchingHierarchyNotes: Comments about the feature matching hierarchy.

FeatureMatchingResultTypeNotes: Comments about the feature matching results.

FeatureMatchingResultCount: Indicates the number of features that the addressed matched. This will usually be just 1.

FeatureMatchingResultTypeTieBreakingNotes: Comments about the logic used in breaking a tie if the address matched to more than one feature.

TieHandlingStrategyType: The strategy used by the geocoder in breaking a tie if the address matched to more than one feature. Is “Unknown” if there was no tie breaking necessary.

Region Size and Region Size Units: This is the size of the feature that the address matched. Building centroids are “-1”.

naaccrQualCode: This is the NAACCR “GIS Coordinate Quality” code from the *NAACCR Standards for Cancer Registries Volume II, Data Standards and Data Dictionary*. Code indicating the basis of assignment of latitude and longitude coordinates for an individual record from an address. This data item is helpful in identifying cases that were assigned coordinates based on incomplete information, post office boxes, or rural routes. Codes are hierarchical, with lower numbers indicating a better quality coordinate. Spatial analysis of cancer data often requires identifying data records with a high degree of geographic precision. Researchers can use this code as a basis for selecting records with a degree of precision that is appropriate to the study.

- 1 **Building Centroid:** Coordinates derived from local government-maintained address points, which are based on property parcel locations, not interpolation over a street segment’s address range
- 2 **GPS:** Coordinates assigned by Global Positioning System (GPS)
- 3 **Exact Parcel Centroid Point:** Coordinates are match of house number and street, and based on property parcel location
- 4 **Address Range Interpolation:** Coordinates are match of house number and street, interpolated over the matching street segment’s address range
- 5 **Street Intersection:** Coordinates are street intersections
- 6 **Street Centroid:** Coordinates are at mid-point of street segment (missing or invalid building number)
- 7 **USPS Zip Plus 4 Centroid:** Coordinates are address ZIP code+4 centroid
- 8 **USPS Zip Plus 2 Centroid:** Coordinates are address ZIP code+2 centroid

- 9 **Manual Lookup:** Coordinates were obtained manually by looking up a location on a paper or electronic map
- 10 **ZCTA Centroid:** Coordinates are address 5-digit ZIP code centroid
- 11 **PO Box ZIP Centroid:** Coordinates are point ZIP code of Post Office Box or Rural Route
- 12 **City Centroid:** Coordinates are centroid of address city (when address ZIP code is unknown or invalid, and there are multiple ZIP codes for the city)
- 13 **County Centroid:** Coordinates are centroid of county
- 98 **State Centroid:** Latitude and longitude are assigned, but coordinate quality is unknown
- 99 **Unmatchable:** Latitude and longitude are not assigned, but geocoding was attempted; unable to assign coordinates based on available information
- Blank **Missing:** GIS Coordinate Quality not coded

naaccrQualType: This is the same as naaccrQualCode, except the code is a number instead of a 2-digit character, and the codes start with 1 instead of 0 for the hierarchy rankings.

- 1 Coordinates derived from local government-maintained address points, which are based on property parcel locations, not interpolation over a street segment's address range
- 2 Coordinates assigned by Global Positioning System (GPS)
- 3 Coordinates are match of house number and street, and based on property parcel location
- 4 Coordinates are match of house number and street, interpolated over the matching street segment's address range
- 5 Coordinates are street intersections
- 6 Coordinates are at mid-point of street segment (missing or invalid building number)
- 7 Coordinates are address ZIP code+4 centroid
- 8 Coordinates are address ZIP code+2 centroid
- 9 Coordinates were obtained manually by looking up a location on a paper or electronic map
- 10 Coordinates are address 5-digit ZIP code centroid
- 11 Coordinates are point ZIP code of Post Office Box or Rural Route
- 12 Coordinates are centroid of address city (when address ZIP code is unknown or invalid, and there are multiple ZIP codes for the city)
- 13 Coordinates are centroid of county
- 0 Latitude and longitude are assigned, but coordinate quality is unknown Blank GIS Coordinate Quality not coded

Micro Match Status - Match/Review/Non-Match status assigned for micro analysis usability.

NOTE: This is a new field from July 2017. Cases in the "Review" category require manual review prior to using the data for geospatial analysis of small areas or area-based measures smaller than counties. These "Review" cases should be run through the NAACCR Geocoder as a separate batch run using "Exhaustive Search." In Fall 2017, NAACCR will make an interface available to assist users in manually selecting among multiple potential addresses (or label as "Non-Match").

OutputGeocodes - Field populated with JSON string containing all the valid geocodes returned for an address if "Exhaustive Search" is selected. A user will need to parse out the multiple addresses.

NOTE: NAACCR is in the process of developing an interface that will handle these multiple addresses and allow a user to sort, filter, review, and manually select (or disregard) the proper addresses. This interface will be available in the Fall of 2017.

3) Census Values Data Items

Most of these data items are census geography codes.

CensusBlock: Census block

CensusBlockGroup: Census blockgroup

CensusTract: Census tract

CensusCountyFips: Census county FIPS (Federal Information Processing Standard) code.

CensusStateFips: Census state FIPS code.

CensusCbsaFips: Census Core-Based Statistical Area.

CensusCbsaMicro: Census Micropolitan Statistical Area. A “Core Based Statistical Area (CBSA) is a collective term for both metropolitan and micropolitan areas.

CensusMcdFips: Census Minor Civil Division.

CensusMetDivFips: Census Metropolitan Division of a Metropolitan Statistical Area.

CensusMsaFips: Census Metropolitan Statistical Area. A “Core Based Statistical Area (CBSA) is a collective term for both metropolitan and micropolitan areas.

CensusPlaceFips: FIPS code for an incorporated place or a census designated place (CDP). Essentially this represents a town.

CensusYear: The census year chosen by the user. The geography of the census units for this particular year is used by the geocoder in assigning census units for the particular address coordinate. The user can choose 1990, 2000, or 2010.

naaccrCertCode & naaccrCertType:

This is the NAACCR “Census Tract Certainty” code from the *NAACCR Standards for Cancer Registries Volume II, Data Standards and Data Dictionary*. Code indicating basis of assignment of census tract for an individual record. Helpful in identifying cases tracted from incomplete information or P.O. Box.

- 1 Residence Street Address: Census tract based on complete and valid street address of residence
 - 2 Residence ZIP Plus4: Census tract based on residence ZIP + 4
 - 3 Residence ZIP Plus2: Census tract based on residence ZIP + 2
 - 4 Residence ZIP: Census tract based on residence ZIP code only
 - 5 PO Box ZIP: Census tract based on ZIP code of P.O. Box
 - 6 Residence City or ZIP with One Census Tract: Census tract/BNA based on residence city where city has only one census tract, or based on residence ZIP code where ZIP code has only one census tract
 - 9 Missing: Not assigned, geocoding attempted. Note that the geocoder will assign this code when the address was geocoded to the city centroid and the city has more than one census tract (usually because the address was a PO Box and the ZIP code is specifically for PO Box addresses and does not have a corresponding ZCTA) or to the state centroid (usually because the town name is not recognized).
- Blank Not Attempted: Not assigned, geocoding not attempted

4) Matched Address Data Items

These data items are the parsed elements of the matched address. Note: not all locations will have every data type

MNumber: House number.

NumberFractional: Fractional element of the house address. Also includes letters, such as “A” in “100 A MAIN ST”, but in many situations that letter is actually an apartment designation.

MPreDirectional: Direction (N, S, E, W) occurring before the street name.

MName: Street name.

MSuffix: Street name suffix (DR, ST, AVE, BLVD, HWY, etc.)

MPostDirectional: Directions (N, S, E, W) occurring after the street name. **MSuiteType:** Unit type (APT, UNIT, RM, SPACE)

MSuiteNumber: Unit number.

MPostOfficeBoxType: PO Box type.

MPostOfficeBoxNumber: PO Box number. **MCity:** Town or city.

MState: State.

MZip: ZIP code, first 5 digits.

MZipPlus1: ZIP +1 (if only 1 digit). **MZipPlus2:** ZIP +2 (if only 2 digits). **MZipPlus3:** ZIP +3 (if only 3 digits).

MZipPlus4: ZIP +4 (if 4 digits – this may be the only column populated of the “ZipPlus” group).

MZipPlus5: ZIP +5 (if 5 digits).

MPreQualifier: Street name qualifier prefix (OLD, NEW).

MPostQualifier: Street name qualifier suffix (SCENIC as in VILLAGE SCENIC PKWY, LOOP as in ABBOTT LOOP RD).

MConsolidatedCity: Consolidated city.

MMinorCivilDivision: Minor Civil Division.

MCountySubRegion: The entire street address is placed in this field.

MCounty: County.

MPreType: Street name prefix that indicates the street type (RUE as in RUE DE LA PAIX, VIA as in VIA BALBOA RD).

MPreArticle: Street name prefix article (LA as in LA TOUCHE BLVD, DE as in DE ARMOUN RD).

MPostArticle: Street name suffix article.)

5) Parsed Address Data Items

These data items are the parsed elements of the matched address.

PNumber: House number.

PNumberFractional: Fractional element of the house address. Also includes letters, such as “A” in “100 A MAIN ST”, but in many situations that letter is actually an apartment designation.

PPreDirectional: Direction (N, S, E, W) occurring before the street name.

PName: Street name.

PSuffix: Street name suffix (DR, ST, AVE, BLVD, HWY, etc.) **PPostDirectional:** Directions (N, S, E, W) occurring after the street name. **PSuiteType:** Unit type (APT, UNIT, RM, SPACE)

PSuiteNumber: Unit number.

PPostOfficeBoxType: PO Box type. **PPostOfficeBoxNumber:** PO Box number. **PCity:** Town or city.

PState: State.

PZip: ZIP code, first 5 digits.

PZipPlus1: ZIP +1 (if only 1 digit). **PZipPlus2:** ZIP +2 (if only 2 digits). **PZipPlus3:** ZIP +3 (if only 3 digits).

PZipPlus4: ZIP +4 (if 4 digits – this may be the only column populated of the “ZipPlus” group).

PZipPlus5: ZIP +5 (if 5 digits).

PPreQualifier: Street name qualifier prefix (OLD, NEW).

PPostQualifier: Street name qualifier suffix (SCENIC as in VILLAGE SCENIC PKWY, LOOP as in ABBOTT LOOP RD).

PConsolidatedCity: Consolidated city.

PMinorCivilDivision: Minor Civil Division.

PCountySubRegion: County Sub-region.

PCounty: County.

PPreType: Street name prefix that indicates the street type (RUE as in RUE DE LA PAIX, VIA as in VIA BALBOA RD).

PPreArticle: Street name prefix article (LA as in LA TOUCHE BLVD, DE as in DE ARMOUN RD).

PPostArticle: Street name suffix article.

6) Feature Address Data Items

These data items are the parsed elements of the feature address to which the original address was geocoded.

These data items are the parsed elements of the matched address.

FNumber: House number. This is a range of numbers if the GeocodeQualityType is “address range interpolation”.

FNumberFractional: Fractional element of the house address.

FPreDirectional: Direction (N, S, E, W) occurring before the street name.

FName: Street name.

FSuffix: Street name suffix (DR, ST, AVE, BLVD, HWY, etc.)

FPostDirectional: Directions (N, S, E, W) occurring after the street name. **FSuiteType**: Unit type (APT, UNIT, RM, SPACE)

FSuiteNumber: Unit number.

FCity: Town or city.

FState: State.

FZip: ZIP code, first 5 digits.

FZipPlus1: ZIP +1 (if only 1 digit).

FZipPlus2: ZIP +2 (if only 2 digits).

FZipPlus3: ZIP +3 (if only 3 digits).

FZipPlus4: ZIP +4 (if 4 digits – this may be the only column populated of the “ZipPlus” group).

FZipPlus5: ZIP +5 (if 5 digits).

FPreQualifier: Street name qualifier prefix (OLD, NEW).

FPostQualifier: Street name qualifier suffix (SCENIC as in VILLAGE SCENIC PKWY, LOOP as in ABBOTT LOOP RD).

FConsolidatedCity: Consolidated city.

FMinorCivilDivision: Minor Civil Division.

FCountySubRegion: County Sub-region.

FCounty: County.

FArea and **FAreaType**: This is the size of the feature that the addressed matched. Building centroids are “-1”. This is the same as **RegionSize** and **RegionSizeUnits**.

FGeometry: Feature geometry, feature available only in desktop version. The geometry returned is in Geography Markup Language (GML). Information about GML available here: <http://www.opengeospatial.org/standards/gml>

FSource: Data source for the feature. A list of the sources and the primary ID field in parenthesis (see definition below) follows:

- SOURCE_BOUNDARY_SOLUTIONS_PARCEL_CENTROIDS (UniqueId), where feature is a parcel
- SOURCE_CENSUS_2010_PLACES (placeFp10), where feature is a city
- SOURCE_CENSUS_2010_STATES (stateFp10), where feature is a state
- SOURCE_CENSUS_2010_ZCTAS (zcta5ce10), where feature is a ZCTA
- SOURCE_CENSUS_TIGERLINES_2010 (tlid), where feature is a street segment

- SOURCE_NAVTEQ_ADDRESSPOINTS_2012 (POINT_ADDRESS_ID), where feature is a building centroid
- SOURCE_NAVTEQ_STREETS_2008 (link_id), where feature is a street segment
- SOURCE_NAVTEQ_STREETS_2012 (link_id), where feature is a street segment

FVintage: Year of the data source.

FPostOfficeBoxType: PO Box type.

FPostOfficeBoxNumber: PO Box number.

FPrimaryIdField: The unique ID field in the feature source database for the feature (listed above).

FPrimaryIdValue: The value in the unique ID field in the feature source database for the feature.

FPreType: Street name prefix that indicates the street type (RUE as in RUE DE LA PAIX, VIA as in VIA BALBOA RD).

FPreArticle: Street name prefix article (LA as in LA TOUCHE BLVD, DE as in DE ARMOUN RD).

FPostArticle: Street name suffix article.

FGeometrySRID: Available with desktop version only. Feature geometry spatial reference identifier. SRIDs are used to uniquely identify the coordinate systems used to define columns of spatial data in a database. Whether it returns a value or not, the coordinates are based on the NAD83 datum which is what the Census Bureau uses for its Tiger files and that SRID is 4269.