

## OSDH Geocoding Standards

### Executive Summary:

The *OSDH GIS Advisory Committee Standards and Policies Workgroup* has developed a set of geocoding standards for GIS software users to apply when geocoding data. Use of the standards will ensure that geocoding is reliable and comparable across agency datasets. The standards will also ensure that data owners and data users have the most information possible about the procedures involved in geocoding and about the quality of the geocoding results. The standards are based on U.S. Postal Service address standards and a compilation of information brought to the *Standards and Policies Workgroup* from several programs at OSDH.

The *GIS Advisory Committee* recommends that GIS users follow the standards when possible and when access to data and resources is available. If the standards cannot be followed because staff do not have access to the appropriate tools or datasets, they should contact the GIS Coordinator for assistance.

An accompanying set of geocoding procedures is provided in the document entitled *OSDH Geocoding Procedures*.

### Definitions:

*Geocoding* is a GIS operation for converting street addresses into spatial data that can be displayed as features on a map, usually by referencing address information from a street segment data layer.

*Address preparation* is a process in which address fields are prepared for the address verification through the following steps:

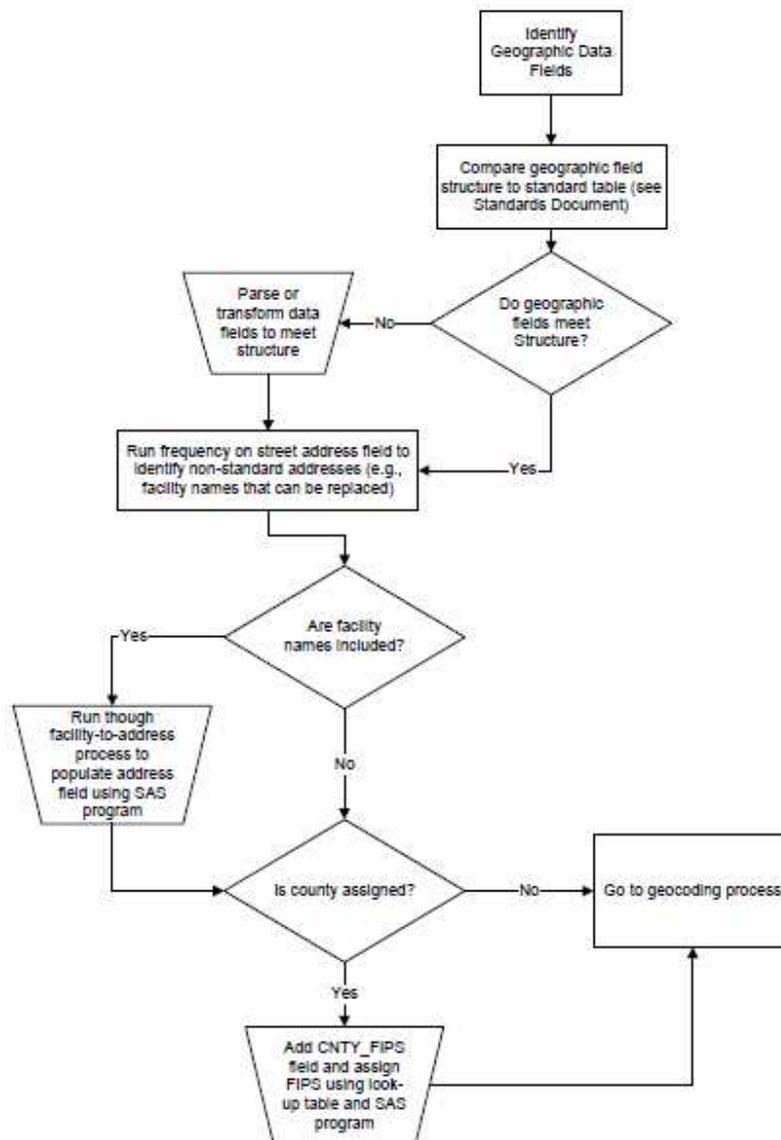
- Parsing address elements into separate fields,
- Replacing facility, building, location or personal name with address.
- Assigning 5-digit FIPS code for county if county is available.

*Address verification* is a process in which addresses are standardized and verified through the following steps:

- Standardizing text and replacing abbreviations
- Assigning street type if possible
- Verifying existence of address
- Changing street type and direction
- Updating of assigned zip+4
- Assigning 5-digit zip code

### Standards for Address Preparation:

Address preparation should follow the *GIS Data Preparation Instructions* as shown in the following diagram.



- If facility names need to be replaced, see Appendix A for sample SAS program
- If 5-digit FIPS code for county needs to be assigned, see Appendix B for county code assignment instructions

Addresses shall be prepared for verification using the following standards:

1. Addresses should contain the following standard address elements in individual fields with descriptive column headings. Address fields should contain the address only, not facility, building, location or personal name.

Field Name	Address Element	Example	Special Instructions
AddLn1	House #	4343	
	Street Prefix Directional	N, E, S, W, NW, SW, etc	Abbreviate
	Street Name	118 <sup>th</sup> , Pennsylvania	Do not abbreviate, Do include “th”, “st”, “nd”, etc
	Street Type	St, Ave, etc	Abbreviate as per Postal Addressing Standards, Appendix C1
	Street Suffix Directional	East, E	Commonly used in the Tulsa area
AddLn2	Secondary Unit Number	12, 3C	
	Secondary Unit Type	APT, STE, TRLR	Abbreviate as per Postal Addressing Standards, Appendix C2
City	City	Oklahoma City	Write out in full or use standard abbreviations only when necessary
State	State	OK	Use two-letter abbreviation
Zip	Zip code	73170, 73170-2534	4-digit add-on is optional but preferred, separate with a dash
Name	Name (optional)	Grace Living Center	Facility name, building name, personal name
County	County (optional)	Oklahoma, 40109	Do not abbreviate; can substitute 5-digit FIPS code for text; include if collected in data
Country	Country (optional)	Canada, Germany	Write out in full

### Standards for Address Abbreviation:

The USPS Abbreviations have been adopted. Refer to, *“Postal Addressing Standards, Appendix C – Street Abbreviations, July 2008”* for the standards, available online.

### Address Verification Software Standards for Batch Geocoding:

The following bulleted list includes minimum requirements for software that performs address verification to be considered acceptable to meet OSDH geocoding standards. One address verification software currently in use at OSDH is Semaphore Corporation’s ZP4. See Appendix C for instructions for using ZP4.

- Minimum of 90% of records in standard dataset assigned zip+4

- Minimum of X% of records in standard dataset assigned county
- Maximum process time of X minutes to run standard dataset for batch processing
- Best value at competitive pricing
- Licensing and maintenance requirements appropriate for utilization
- Batch processing capability
- User-friendly interface with accessible training materials
- Minimum required input fields to include id, street address, city, state and zip
- Minimum required output fields to include input fields plus standardized and verified street address, verified city, verified state, verified zip+4, county, county FIPS

### Standard Row-level Geocoding Tiers by Geocoding Method including Standards for Geographical Input Used for Geocoding (by Tier):

The geocoding standards include guidelines for assigning a quality code to each row of data to indicate how the address in that row was geocoded or not. The codes can then help a data user know the spatial accuracy of the geocoding per row and per dataset. Knowing the spatial accuracy will help contributed to decisions about how to use the data in spatial analysis and mapping. The following table lists the tier codes, their definitions and other pertinent information required for the geocoding staff and the data users.

Tier	Definition	Geographical Input	Minimum Matching Requirement	Address Locator
1A	Coordinates are obtained using GPS technology at the main building entrance	Not Applicable	Not Applicable	Not Applicable
1B	Coordinates are obtained using GPS technology at the main street entrance (intersection of main driveway and street)	Not Applicable	Not Applicable	Not Applicable
1C	Coordinates are obtained using GPS technology at any other location on the property	Not Applicable	Not Applicable	Not Applicable
2A	Location is geocoded at the point/centroid level using address point/structure centroid	Address point/structure centroid obtained from local county tax assessor or commercial	100%	Pending data

Tier	Definition	Geographical Input	Minimum Matching Requirement	Address Locator
		vendor		
2B	Location is geocoded at the point level using a street frontage point marking	Commercial vendor	100%	Pending data
2C	Location is geocoded at the centroid level using a parcel centroid	Parcel centroids obtained from local county tax assessor or commercial vendor	100%	Pending data
3A	Location is geocoded at the street address level by a manual match using personal communication with the facility	Online resources or map of roads data (e.g., Google, Bing, MS Virtual Earth, etc.)	Not Applicable	Not Applicable
3B	Location is geocoded at the interpolated street address level using NAVTEQ or TeleAtlas roads data with alternate street names	Commercial roads data	72%	NAVTEQ_Address_3B
3C	Location is geocoded at the interpolated street address level using roads data from local sources (i.e. ACOG or INCOG datasets)	ACOG roads data, INCOG roads data, or other local data source	72%	INCOG_Address_3C, ACOG_Address_3C
3D	Location is geocoded at the interpolated street address level using TIGER roads data	TIGER roads data	72%	Pending data
3E	Location is geocoded at the interpolated street address level by a manual match using online resources	Online resources (e.g., Google, Bing, MS Virtual Earth, etc.)	Not Applicable	Not Applicable
3F	Location is geocoded to the street intersection	Commercial roads data or online resources	100%	NAVTEQ_Address_3B
3G	Location is geocoded to the mid-point of the street segment with city	Commercial roads data or online resources	100%	Pending development

Tier	Definition	Geographical Input	Minimum Matching Requirement	Address Locator
	and zip			
4A	Location is geocoded at the centroid level using a Zip+4 centroid	Commercial Zip+4 centroids	100%	NAVTEQ_Zip4_4A
4B	Location is geocoded at the centroid level using a Zip+2 centroid	Commercial Zip+4 centroids	100%	NAVTEQ_Zip2_4B
4C	Location is geocoded at the 5 digit zipcode level using a combination of the 5 digit zipcode and city	Commercial 5-digit zip code data and tax commission municipal boundaries	100%	NAVTEQ_Zip5_Munbound_4C
4D	Location is geocoded at the 5 digit zipcode level using the 5 digit zipcode centroid only for street addresses or rural routes	Commercial 5-digit zip code data	100%	NAVTEQ_Zip5_4D
4E	Location is geocoded at the 5 digit zipcode level using the 5 digit zipcode centroid of a PO Box or Organization	Commercial 5-digit postal zip code data	100%	NAVTEQ_Zip5_4E
5A	Location is geocoded at the city level using a derived centroid from tax data	Centroid of city polygon from tax assessor	100%	Munbound_5A
5B	Location is geocoded at the city level using a city point	Commercial city centroid data	100%	NAVTEQ_City_5B
6A	Latitude and longitude are assigned, but coordinate quality is unknown	Various data	Not Applicable	Not Applicable
6B	Latitude and longitude are not assigned, but geocoding was attempted; unable to assign coordinates based on available information	Not Applicable	Not Applicable	Not Applicable

## Standards for Including Dataset Geocoding Quality to Metadata

Metadata should be completed for all GIS datasets following the *OSDH Metadata Template Usage Guidelines*. Metadata should include information about the quality of geocoding for the dataset. The abstract should include language such as, “The data were geocoded at Tiers 3B and 3D (see OSDH Geocoding Standards) by address using the 2009 NAVTEQ NAVSTREETS data and online resources. The address used for geocoding was the physical address of the facility as reported by the hospital.”

Relevant tiers should be defined in the attribute metadata. For example, the GCTIER should be defined as follows with some modifications made as appropriate for the dataset. “GCTIER - 3B, 3D. Indicates the tier at which geocoding was completed (see OSDH Geocoding Standards). Tier 3B indicates location was geocoded at the interpolated street address level using NAVTEQ roads data with alternate street names. Tier 3D indicates the location was geocoded at the interpolated street address level by a manual match to NAVTEQ roads using online resources.”

## Standards for Storage of Geocoded Data

After the geocoding process is complete, formats for storage of the data will depend on the type of data and program requirements.

If the data represent person-centric residence addresses:

- The attribute table (including lat/long coordinates and other geographic fields – see below) should be stored in the location most accessed by users of the data, depending on program requirements. It can be stored in any tabular format accessible by data users.
- The shapefile that contains all output fields from geocoding should be stored locally by the geocoder/data owner.
- For the *person-centric attribute table*, the following table lists the fields to keep and to drop. The fields indicated to keep should be the most fields kept in the attribute table. Fewer may be kept if requested by user.

Field Name	Description	Origin	Keep	Drop
FID	System generated feature id	ARC		✓
Shape	Type of shape (point, polygon, etc.)	ARC		✓
Status	Matched, Unmatched	ARC		✓
Score	Match score	ARC		✓
Match_type	Automatic, manual or pick point match	ARC		✓
Side	Side of street	ARC		✓
X (rename as longitude)	X coordinate	ARC	✓	
Y (rename as latitude)	Y coordinate	ARC	✓	
Match_addr	The address matched to	ARC	✓	
ARC_street	The street address according to ARC	ARC		✓
ARC_city	The city according to ARC	ARC		✓
ARC_ZIP	The ZIP according to ARC	ARC		✓
Record_id	Program record id	Original data	✓	
Street	Program street	Original data	✓	

City	Program city	Original data	✓	
State	Program state	Original data	✓	
Zip	Program zip	Original data	✓	
New_street	Cleaned street	Cleaning	✓	
Ad_zp4	Street address according to ZP4	ZP4 address verification		✓
City_zp4	City according to ZP4	ZP4 address verification		✓
St_zp4	State according to ZP4	ZP4 address verification		✓
Zip_zp4	Zip+4 according to ZP4	ZP4 address verification		✓
County_zp4	County according to ZP4	ZP4 address verification		✓
Fips_zp4	County 2-digit FIPS code according to ZP4	ZP4 address verification		✓
LACS	ZP4 Indicator if converted from RR to e911	ZP4 address verification		✓
GCTier	Tier number	Geocoder assigned	✓	
CntyAssn	How the county was assigned	Geocoder assigned		✓
County	County Name	Geocoder assigned	✓	
Cnty_FIPS	5-digit FIPS	Geocoder assigned	✓	
CTID	In state census tract #	Geocoder assigned	✓	
Shape_leng	System generated feature length value	ARC		✓
Shape_area	System generated feature area value	ARC		✓

If the data represent facility locations:

- The shapefile can be submitted to the GIS Coordinator to be considered for storage in the OSDH Geodatabase if the data would be of interest to more than one GIS user.
- The attribute table (including lat/long coordinates and geocoding output fields) will be saved in a backup location by the GIS Coordinator.
- The shapefile should only be stored locally if not stored in the Geodatabase.
- Fields kept for the *facility location shape file* will be dictated by the geodatabase schema. All output fields will be kept for the *facility location attribute table* in a backup location.

Each geocoded dataset should be stored with a *Geocoding Results Report*, stored as a Word document. See Appendix D for the reporting tool template.

### Document Revision History

- Revised on 10-27-09 – added content
- Revised on 2-16-2010 – added content and made revisions, added appendices

## Appendix A – Facility-to-Address Sample Program for SAS

The following is just a small part of the full code. For a digital copy of the code, please contact the GIS Coordinator.

```

/* cleaning up the address field */

/* removing characters and extra spaces at the beginning of the field and making all the text upper case */
data facility;
  set dischcln.t_discharge04_clean;
  if substr(address,1,1) in ('#','%') then address=substr(address,2);
  else if substr(address,1,2) = 'MR' then address=trim(substr(address,3));
  else if substr(address,1,22)='*****DECEASED*****' then address=substr(address,23);
  else if substr(address,1,20)='*****DECEASED*****' then address=substr(address,21);
  else if address in ('HOMELESS','!!RETURNED MAIL!!!','HOMELESS','U','UK','UN','UNK','UNK NURSING HOME',
    'UNKNOWN','UNKNOWN ADDRESS','UNKOWN','UPDATE ADDRESS','X','XY','NEED ADDRESS','BAD ADDRESS','NO GOOD
ADDRESS','PRE INS ONLY',
    'RETURN MAIL','NOT AVAILABLE','RETURNED MAIL','BAD ADDRESS*****','BAD ADDR','***RETURN MAIL/ADDRESS
INSE***','*** BAD ADDRESS ***')
    then address=' ';
  else address=trim(uppercase(address));

  if city in ('LAMTZENHAUSEN, GERMANY','AAAAAAAAAAAAAAAA') then city=' ';
  else city=trim(uppercase(city));
run;
/* running the facility to address program to determine replace facility names with addresses */
data dischcln.t_discharge04_clean;
  set facility;
  if not ('0'<=substr(address,1,1)<='9') and substr(address,1,2) not in ('RT','PO','RR','HC','R1','R2','BX','R5','R6')
    and substr(address,1,3) not in ('BOX','P O','P.O','NBU','R 1','R.T','R R','HWY','H C','H.C','LOT','PSC','STAR')
    and substr(address,1,4) not in ('RUAL','BLDG')
    and substr(address,1,5) not in ('ROUTE','P. O.','P O ','RURAL') and substr(address,1,6) not in ('DRAWER','GENERA','R ROUT')
    and address ^= ' ' then do;
  if city = 'ADA' then do;
    if index(address,'ROLLING HILLS') >0 then newstreet = '1000 ROLLING HILLS LANE';
    else if index(address,'STONEGATE') >0 then newstreet = '130 E 6TH STREET';
    else if index(address,'MCCALL') >0 then newstreet = '13546 COUNTY ROAD 3600';
    else if index(address,'HEARTLAND') >0 or index(address,'HARTLAND')>0 then newstreet = '1501 NORTH MONTE VISTA';
    else if index(address,'BALLARD') >0 or index(address,'BALLARD')>0 then newstreet = '201 WEST 5TH STREET';
    else if index(address,'ENCOMPASS') >0 then newstreet = '2600 ARLINGTON ST # B';
    else if index(address,'ADA SENIOR') >0 then newstreet = '301 E KINGS ROAD';
    else if index(address,'BAPTIST') >0 then newstreet = '3501 NORTH OAKRIDGE BOULEVARD';
    else if index(address,'ADA RESIDENTIAL') >0 or index(address,'WOODLAND')>0 then newstreet = '626 W 15TH ST';
    else if index(address,'STERLING') >0 then newstreet = '801 SOUTH STADIUM DRIVE';
    else if index(address,'JAN FRANCES') >0 or index(address,'JFCC')>0 or index(address,'JAN FRANCIS')>0 then newstreet = '815
NORTH COUNTRY CLUB ROAD';
    else if index(address,'SLEEPY HOLLOW') >0 then newstreet = '827 RAILHEAD DR';
    else if index(address,'ADA RETIREMENT') >0 then newstreet = '931 NORTH COUNTRY CLUB ROAD';
    else newstreet = " ";
  end;
  else if city = 'ALLEN' then do;
    if index(address,'WOODLAND HILLS') >0 then newstreet = '200 N EASTON ST';
    else newstreet = ' ';
  end;
  else if city = 'ALTUS' then do;
    if index(address,'JACKSON COUNTY') >0 then newstreet = '1200 E PECAN ST';
    else if index(address,'TAMARACK') >0 then newstreet = '1224 E TAMARACK RD';
    else if index(address,'ENGLISH VILLAGE') >0 then newstreet = '1515 CANTERBURY BLVD';
  end;

```

```

    else if index(address,'SOUTHWEST ADULT DAY CARE') >0 then newstreet = '2208 ENTERPRISE DR';
    else if index(address,'PLANTATION ') >0 or index(address,'PLANTATTION ') >0 or index(address,'GRACE LIVING ') >0 then
newstreet = '2610 CEDAR CREEK DR';
    else if index(address,'STATE OF OK PAMELA GRAN') >0 then newstreet = '308 W BROADWAY ST';
    else newstreet = ' ';
end;
else if city = 'ALVA' then do;
    if index(address,'BILL JOHNSON') >0 or index(address,'BJCC') >0 then newstreet = '1856 E FLYNN STREET';
    else if index(address,'SHARE MEDICAL') >0 then newstreet = '730 SHARE DRIVE';
    else if index(address,'SHARE CONVALESCENT') >0 or index(address,'SHARE CONVALESENT')>0 then newstreet = '800 SHARE
DRIVE';
    else if index(address,'BEADLES') >0 then newstreet = '916 NOBLE ST';
    else newstreet = ' ';
end;
else if city = 'ANADARKO' then do;
    if index(address,'SILVERCREST') >0 or index(address,'SILVER CREST') >0 then newstreet = '300 W WASHINGTON AVE';
    else if index(address,'GLENHAVEN') >0 then newstreet = '3003 IOWA';
    else newstreet = ' ';
end;
else if city = 'ANTLERS' then do;
    if index(address,'SOONER APT') >0 then newstreet = '1100 SOONER DR';
    else if index(address,'CHOCTAW NATION') >0 then newstreet = '400 SOUTHWEST O STREET';
    else if index(address,'ANTLERS') >0 then newstreet = '511 EAST MAIN';
    else newstreet = ' ';
end;
else if city = 'ARDMORE' then do;
    if index(address,'VETERAN') >0 then newstreet = '1015 S COMMERCE ST';
    else if index(address,'WHISPERING OAKS') >0 then newstreet = '111 13TH AVE NW';
    else if index(address,'VILLAGE LODGE OF ARDMORE VILLAGE') >0 then newstreet = '1310 KNOX RD';
    else if index(address,'WOODVIEW') >0 then newstreet = '1630 3RD AVE NE';
    else if index(address,'SUITES AT ELMBROOK') >0 then newstreet = '1711 9TH AVE NW';
    else if index(address,'ELMBROOK HOME') >0 then newstreet = '1811 9TH AVE NW';
    else if index(address,'SUNSHINE ADULT') >0 then newstreet = '945 10TH AVE SE';
    else if index(address,'ADULT DAY SERVICES OF SOUTHERN') >0 then newstreet = '1902 SHENANDOAH DR';
    else if index(address,'HEARTLAND PLAZA') >0 then newstreet = '2215 4TH AVE NW';
    else if index(address,'WESTERN HILLS') >0 then newstreet = '402 PAWNEE ST NW';
    else if index(address,'VILLAGE LODGE') >0 then newstreet = '5361 BRISTOL PIKE';
    else if index(address,'LAKELAND MANOR') >0 then newstreet = '604 LAKE MURRAY DRIVE';
    else if index(address,'SOUTHBROOK HEALTHCARE') >0 then newstreet = '832 ISABEL ST';
    else if index(address,'SUNSHINE ADULT') >0 then newstreet = '945 10TH AVE SE';
    else newstreet = ' ';
end;
else if city = 'ARKOMA' then do;
    if index(address,'MEDI-HOME') >0 or index(address,'MEDI HOME') then newstreet = '1008 ARKANSAS STREET';
    else newstreet = ' ';
end;
else if city = 'ATOKA' then do;
    if index(address,'GRASSEY') >0 then newstreet = '1099 GRASSEY LAKE RD';
    else if index(address,'SOUTHWIND') >0 then newstreet = '125 WALKER CIRCLE';
    else if index(address,'ATOKA MANOR') >0 then newstreet = '1500 S VIRGINIA AVE';
    else if index(address,'HMCC') >0 or index(address,'MCLEOD ') >0 then newstreet = '1970 E WHIPPOORWILL LANE';
    else if index(address,'WEST CEDAR CIR') >0 then newstreet = '71 W CEDAR CIR';
    else newstreet = ' ';
end;
else if city = 'BARNSDALL' then do;
    if index(address,'BARNSDALL NURSING HOME') >0 then newstreet = '411 SOUTH CHESTNUT';
    else newstreet = ' ';
end;

```

## Appendix B – County FIPS Code Assignment Instructions

Discharge Data Geocoding example:

### 1. Address replacement for facility names

- Old Method:
  - Convert access database to tab-delimited text file (pad first row values with x's to allow all data to be read into SAS)
  - Import text file into SAS and run through “*Proc SQL Distinct County.sas*” contained in the Geocoding\_Projects > SAS\_Codes folder (code will require some tweaking for file/field names) to identify invalid addresses
  - Export resulting set to a text file and import into Excel
  - Delete addresses that cannot be replaced (i.e. c/o John Smith, no address, homeless, etc), leaving only facility names
  - Create a regular expression to be used to convert the Excel table into SQL code to update the “NewStreet” field
  - Open “Update New Street Column.sas” and replace the code with the new code from the regular expression (file/field names will need to be tweaked also); this will update the NewStreet field with the addresses of the facilities
- New Method: run the “*fac\_to\_address pgm.sas*” that Becki wrote

### 2. Address verification/correction in ZP4 software

- Export the data table from SAS to a tab delimited text file
- Import the text file into Excel and save.
- Run the text file through ZP4 on the laptop (following the “*ZP4\_Instructions.docx*”). Also output county name, county code (this is a 3-digit code) and state code from ZP4 for this data.
- Import the output from ZP4 into Excel next to the original data starting at row 2 (because there are no headers on the ZP4 output).
- Do a visual check to make sure the original data and the ZP4 output align correctly (this should be the case as long as the data wasn't sorted after the ZP4 run).
- Add field headers
- Add two fields at the end called GCTIER (format: text) and CntyAssn (text)
- Save the Excel file as a 2007 version

### 3. Geocoding in ArcMap

- a) GCTIER 3B (NAVTEQ):
  - Add the Excel file to ArcMap for geocoding
  - Run first geocoding pass, using the *NAVTEQ\_Address\_3B locator* (ex: save as **Outpt2009Geo3B.shp**)
  - Enter statistics into the “*Geocoding Log*” and populate GCTIER as 3B for matched and tied records only
  - Select unmatched records from the shapefile and export as a dbf table, Outpt2009Geo3B\_UM.dbf
- b) GCTIER 3B (ESRI):

- Run the unmatched records through the *ESRI\_Street\_Addresses\_US locator* (ex: save as **Outpt2009Geo3Besri.shp**)
  - Enter statistics into the “*Geocoding Log*” and populate GCTIER as 3B for matched and tied records only
  - Select unmatched records from the shapefile and export as a dbf table, Outpt2009Geo3Besri\_UM.dbf
  - Select records from the shapefile from this step where State\_ZP4 is not equal to OK and Status = ‘M’
  - Export selected records as **Outpt2009GeoOOS\_M.shp**
- c) GCTIER 3C (INCOG):
- Geocode the Outpt2009Geo3Besri\_UM.dbf table using the *INCOG\_Address\_3C locator* (ex: save as **Outpt2009Geo3C\_INCOG.shp**)
  - Enter statistics into the “*Geocoding Log*” and populate GCTIER as 3C for matched records only
  - Select unmatched records from the shapefile and export as a dbf table, Outpt2009Geo3C\_INCOG\_UM.dbf
- d) GCTIER 3C (ACOG):
- Geocode the Outpt2009Geo3C\_INCOG\_UM.dbf table using the *ACOG\_Address\_3C locator* (ex: save as **Outpt2009Geo3C\_ACOG.shp**)
  - Enter statistics into the “*Geocoding Log*” and populate GCTIER as 3C for matched records only
  - Select unmatched records from the shapefile and export as a dbf table, Outpt2009Geo3C\_ACOG\_UM.dbf
- e) GCTIER 4A:
- Geocode the unmatched records from step d using the *NAVTEQ\_Zip4\_4A locator*, (ex: save as **Outpt2009Geo4A.shp**)
  - Enter statistics into the “*Geocoding Log*” and populate GCTIER as 4A for matched records only
  - Select unmatched records from the shapefile and export as a dbf table, Outpt2009Geo4A\_UM.dbf
- f) GCTIER 4D:
- Geocode the unmatched records from step e using the *NAVTEQ\_Zip5\_4D\_NotOverlap locator*, (ex: save as **Outpt2009Geo4D.shp**)
  - Enter statistics into the “*Geocoding Log*” and populate GCTIER as 4D for matched records only
  - Select unmatched records from the shapefile and export as a dbf table, Outpt2009Geo4D\_UM.dbf
- g) GCTIER 4E:
- Geocode the unmatched records from step f using the *NAVTEQ\_Zip5\_4E locator*, (ex: save as **Outpt2009Geo4E.shp**)

- Enter statistics into the “*Geocoding Log*” and populate GCTIER as 4E for matched records only
  - Select unmatched records from the shapefile and export as a dbf table, **Outpt2009Geo4E\_UM.dbf**
- h) GCTIER 4D5A:
- Add a new field to the unmatched records output from step g and name the field CityZip (format: text, length: 50)
  - Populate the CityZip field by appending the City\_ZP4 and Zip5\_ZP4 values together
  - Geocode the unmatched records from step g using the *NAVTEQ\_Zip5\_City\_NotOverlap locator* and the CityZip field (this uses polygons that were created based on unique combinations of city boundaries and zip code boundaries; only the polygons that do not overlap county boundaries were retained), (ex: save as **Outpt2009Geo4D5A.shp**)
  - Enter statistics into the “*Geocoding Log*” and populate GCTIER as 4D5A for matched records only
  - Select unmatched records from the shapefile and export as a dbf table, **Outpt2009Geo4D5A\_UM.dbf**
4. Assigning County names, FIPS codes and census tract IDs
- County/Census tract assignment:
    - Open the Identity tool from ArcToolbox (Analysis Tools > Overlay > Identity)
      - For the Input Features enter the shapefile that was output from step a (the one in **bold**)
      - For the Identity Features, input the *NAV\_CensusTracts* feature class from the *Geocoding\_NAVTEQ* dataset in the geodatabase
      - Name the output feature class as **Outpt2009Geo3B\_CntyCT.shp**
      - Leave all other default values
      - Click OK to run the tool. This assigns the county name (CO\_NAME), 5-digit county FIPS code (FIPSTCO), and the census tract ID (ID) to each geocoded point.
      - *Note:* The number of records in the Identity output should equal the sum of matched and tied records from the shapefile created in step a. The identity output could have slightly more records, though (by just a few). This would happen if a geocoded point intersects the boundary between census tracts; when that happens, the tool returns one record for each intersection that the point participates in. These duplicate records should be dealt with later in SAS with a PROC FREQ step.
      - Repeat the above steps for shapefiles created in steps b-e. Check the count of records against the matched and tied records in the original shapefile after each iteration.
    - After each Identity step is completed, populate the CntyAssn field with the following value for all records “Assigned by point/county intersection”
  - County assignment
    - Open the Identity tool
      - For the Input Features enter the shapefile that was output from step f

- For the Identity Features, input the *NAV\_Counties* feature class from the *Geocoding\_NAVTEQ* dataset in the geodatabase
- Name the output feature class as **Outpt2009Geo4D\_Cnty.shp**
- Leave all other default values
- Click OK to run the tool. This assigns the county name (COUNTY) and 5-digit county FIPS code (CNTYFIPS) to each geocoded point. You will not assign census tract for these points because the geocoding is not accurate enough to assign census tract. You might find that some records are duplicated through this process, too.
- Repeat the above steps for shapefiles created in steps f-h
- After each Identity step is completed, populate the CntyAssn field with the following value for all records “Assigned by point/county intersection”

#### 5. Exporting data from ArcMap to work with in Excel

- Based on the criteria in the following table, you will export the GIS data as text files that will be imported into Excel.

Original file	Originating step	Selection criteria	Exported file (save to OutpatientFinal folder)
Outpt2009Geo3B_CntyCT.shp	4	None	Outpt2009Geo3Bxxxx.txt (xxxx = the number of records being exported)
Outpt2009Geo3Besri_CntyCT.shp	4	None	Outpt2009Geo3Besrixxxx.txt
Outpt2009GeoOOS_M.shp	3,b	None	Outpt2009GeoOOSxxxx.txt
Outpt2009Geo3C_INCOG_CntyCT.shp	4	None	Outpt2009Geo3C_INCOGxxxx.txt
Outpt2009Geo3C_ACOG_CntyCT.shp	4	None	Outpt2009Geo3C_ACOGxxxx.txt
Outpt2009Geo4A_CntyCT.shp	4	None	Outpt2009Geo4Axxxx.txt
Outpt2009Geo4D_Cnty.shp	4	None	Outpt2009Geo4Dxxxx.txt
Outpt2009Geo4E_Cnty.shp	4	None	Outpt2009Geo4Exxxxx.txt
Outpt2009Geo4D5A_UM.dbf	3,h	State_ZP4 ≠ OK and State_ZP4 ≠missing	Outpt2009UMOOSxxxx.txt (UM = unmatched, OOS = out of state and xxxx = the number of records being exported)
Outpt2009Geo4D5A_UM.dbf	3,h	County_ZP4 is null, none or invalid (i.e. county that is not an Oklahoma	Outpt2009UM_manualxxxx.txt

Original file	Originating step	Selection criteria	Exported file (save to OutpatientFinal folder)
		county)	
Outpt2009Geo4D5A_UM.dbf	3,h	State_ZP4 = OK and County_ZP4 is a valid OK county	Outpt2009UM_zp4assnxxx.txt

#### 6. Adding data to Excel

- Open the Excel file called “*Template\_Manual County Assign & Append.xlsx*” from the *Geocoding\_Projects* folder.
  - This template has already been set up with one worksheet for each file exported from ArcMap.
  - The first worksheet is there to hold all the data once it is appended together
- Import each of the text files into its appropriate worksheet in the Excel file
- Verify that all data was imported correctly
- Standardize the number of columns, column names and column order across all worksheets (I usually copied the header row from the 3B output and pasted it above the header row in each worksheet, then rearranged each worksheet as needed to meet the header row; this may require some tweaking)

#### 7. Manual county assignment

- Sort the records in the UMmanualxxxx worksheet by Zip5\_ZP4
  - For GCTIER, enter 6B to indicate geocoding was attempted but unsuccessful
  - For records with missing or invalid zip codes, if the county cannot be estimated from the address or city, enter the county as 00000 (indicates unknown) and CntyAssn as “Unassigned, in state”
  - For all other records, if time allows and the number of records with physical address is minimal, an attempt at manually locating the address on Google maps or Bing maps can be undertaken. Bing maps displays the county boundaries to identify county but Google does not.
    - For any records where county is manually assigned by mapping the address online, enter the CntyAssn value as “Assigned manually”
- For all other records, a proportional county assignment method should be used.
  - Open the Excel file called “*Counties and Zips\_Melissa Data.xlsx*” in the *Geocoding\_Projects* folder. This spreadsheet shows the percent of addresses from each zip code in the state that fall in one county or another. The table is sorted by zip code. The data was obtained from melissadata.com in 2009.
  - Use the Melissa Data percentages to randomly assign addresses to a county based on the percent of addresses in each county.

- Example: Zip code 73010 has 100 records in the unmatched worksheet. Melissa Data shows that 73010 crosses Grady and McClain counties with 30.5% of addresses in Grady county and 69.5% of addresses in McClain county. Based on this, you should assign 30 of the records to Grady and 70 to McClain.
    - For all records with county assigned using the proportionate method, populate the CntyAssn field as “Assigned manually by zip/county proportion”
  - For records in the GeoOOSxxxx worksheet, populate County fields with Out of state and 99998 and CntyAssn with “Unassigned, out of state”
  - For records in the UMOOSxxxx worksheet, populate GCTIER as 6B, County fields with Out of state and 99998, and CntyAssn as “Unassigned, out of state”
  - For records in the UMzp4xxxx worksheet, populate GCTIER as 6B and CntyAssn as “Assigned by ZP4 US Postal Database”. Also, ensure that the county codes have 5 digits (add “40” to the front of the 3 digit codes if necessary)
8. Appending data and post-processing in SAS
- Once all the worksheets have the same schema and have values entered for county, countyFIPS, census tract, GCTIER, CntyAssn, etc, the data can be combined into the first worksheet
  - Copy and paste the contents of each worksheet (except the header rows) into the first worksheet
  - Add a dummy row 1 and pad the values with x’s and zeroes so that row will encompass the maximum field length for the entire dataset (so that all data will be read into SAS)
  - Save the big worksheet as a text file and import into SAS
  - Open the code “*Inp2008PostProcessing.sas*” in SAS (from the Geocoding\_Projects > SAS\_Codes folder)
  - Edit the code to work with the appropriate data file and run the various PROC FREQ steps
    - The first PROC FREQ step identifies the records that have duplicate pk\_event numbers so you can deal with those records that were duplicated during the Identity steps. Choose one of the duplicates to retain and delete the others.
    - Other PROC FREQ steps are used to quality assure the data and also to generate the statistics used to populate the geocoding results report (see “*Inpatient 2008 Geocoding Results for Binitha.docx*” in O:\Geocoding\_Projects\Inpatient\Inp2008)
    - There are also some DATA steps used to correct errors (including discrepancies in spelling of the “Mc” counties)
  - Once the SAS coding is complete, export the final dataset as a tab delimited text file (DON’T export as a csv, I did that once and it scrambled the data a little because of commas in the field values)
  - Import the text file into an Access database
    - Keep only the data fields listed in the geocoding results report at the top (other fields should be retained in the GIS files and SAS file but Binitha only needs certain data fields back)
    - Rename fields if necessary

- Run some queries to verify that the data imported correctly
- Provide the Access database and the geocoding results report to Discharge Project Coordinator

## Appendix C - Address Verification with ZP4 Software

“In **batch mode**, ZP4 can process a database of any size, automatically correcting and standardizing every address in the file. ZP4 can process address lists in dBase, FoxPro, Access, Excel, FileMaker, Oracle, SQL Server, fixed-field text, delimited text, and other file formats.” (Semaphore Corporation, [www.semaphore.com](http://www.semaphore.com))

The ZP4 program is used by Health Care Information to produce standardized addresses within tabular datasets to enable geocoding of the records for mapping with GIS software. The following steps give the user a guide to the use of ZP4 in this context.

1. Start with a table of addresses in an Excel file. Typically, the file consists of records, with each record containing at least an identifier, name, street address, second address line, city, state and ZIP code. The first row should be a header record that identifies or names each column, or field.
2. Choose a directory and save the Excel file as a tab delimited text file (\*.txt).
3. Close the new file.
4. Navigate to ZP4 at Start>All Programs>ZP4>ZP4batch (on Katy Rich’s computer or laptop).
5. Select - variable-length delimited.txt file.
6. Select the .txt file from Step 2 to process.
7. Text quoting: select - none.
8. Field separator: select - tab.
9. Click OK.
10. To map the address fields:
  - a. Click and drag **address field labels** into the **field name** column opposite **field contents** from the address table.
  - b. To un-map a field, double-click on it in the **field contents** column.
11. Check - first record is a header.
12. Click OK.
13. Select output fields which will be produced in the output file; for example:
  - a. Address (final)
  - b. City (final)
  - c. State (final)
  - d. ZIP (five-digit)
  - e. ZIP (four-digit add-on)
  - f. ZIP (final)
14. Select text quoting - Double quotes.
15. Select Field Separator – Tab.
16. Check – Create output header record.
17. Click OK.
18. Click Yes to confirm.
19. Name the output file – replace batchout.txt with the file name you want to use (\*.txt).
20. Give the output file a directory address where you want the file to be located.
21. Click – OK.
22. Click – GO. (ZP4 processes the input file and returns the number of records processed.)
23. Click OK and close ZP4 (two tabs).
24. Open the original Excel file.
  - a. Import text from the ZP4 output file, joining the new text to the existing records.

- b. Rename the ZP4 fields as:
- |                            |   |             |
|----------------------------|---|-------------|
| i. Address (final)         | → | Address_ZP4 |
| ii. City (final)           | → | City_ZP4    |
| iii. State (final)         | → | State_ZP4   |
| iv. ZIP (five-digit)       | → | ZIP5_ZP4    |
| v. ZIP (four-digit add-on) | → | ZIP4_ZP4    |
| vi. ZIP (final)            | → | ZIP9_ZP4    |
- c. Set each column as text format (except numerical fields = general format).
- d. Save as a new Excel file (\*\_ZP4.xlsx).
- e. Verify that all address fields are in text format.
- f. Name the worksheet.
- g. Delete any extra (empty) sheets in the file.
- h. Close and save the file.
- i. This is the file that will be processed when the table is geocoded.

**Appendix D – Template Geocoding Results Report**

The following template can be used to report geocoding results to data owners, data users, and others.

**Geocoding Report**

Dataset name: \_\_\_\_\_

Date geocoded: \_\_\_\_\_

Total records: \_\_\_\_\_

<b>Fields Included in Final Table</b>			
NewStreet	Address where facility names were replaced with street addresses		
Ad_zp4	Standardized street address from ZP4 address verification software		
City_zp4	Verified city from ZP4 address verification software		
St_zp4	Verified state from ZP4 address verification software		
Zip_zp4	Verified 5 or 9 digit zip from ZP4 address verification software		
Longitude	Longitude coordinates		
Latitude	Latitude coordinates		
County	Assigned county name		
Cnty_FIPS	5 digit county code (99998 = out of state, 00000 = unknown)		
CTID	11 digit census tract code		
GCTier	Description of how the address was geocoded or not (if unmatched)		
<b>Geocoding Results</b>			
<b>Geocoding Tier</b>	<b>Count</b>	<b>Percent</b>	<b>Target Range*</b>
* Based on previous geocoding results for similar dataset(s)			
<b>County Assignment of Unmatched Records (optional)</b>			
<b>County Assignment Method</b>	<b>Count</b>	<b>Percent</b>	<b>Target Range<sup>+</sup></b>
Assigned by ZP4 US Postal Database			
Assigned manually			
Assigned manually by zip/county proportion			
Unassigned, out of state			
Unassigned, in state			
<sup>+</sup> Based on previous county assignment results for similar dataset(s)			
<b>Census Tract Assignment of Unmatched Records (optional)</b>			

<b>County Assignment Method</b>	<b>Count</b>	<b>Percent</b>	<b>Target Range<sup>+</sup></b>
Assigned by ZP4 US Postal Database			
Assigned manually			
Assigned manually by zip/census tract proportion			
Unassigned, out of state			
Unassigned, in state			
<b>County Assignment Frequencies (optional)</b>			
<b>CountyName</b>	<b>Cnty_FIPS</b>	<b>Count</b>	<b>Percent</b>
Unknown	00000	35	0.01
Adair	40001	2957	0.58
Alfalfa	40003	865	0.17
Atoka	40005	2058	0.40
Beaver	40007	430	0.08
Beckham	40009	3820	0.75
Blaine	40011	1955	0.38
Bryan	40013	6317	1.23
Caddo	40015	4357	0.85
Canadian	40017	12673	2.47
Carter	40019	7831	1.53
Cherokee	40021	5470	1.07
Choctaw	40023	3182	0.62
Cimarron	40025	172	0.03
Cleveland	40027	29227	5.70
Coal	40029	1188	0.23
Comanche	40031	12438	2.43
Cotton	40033	714	0.14
Craig	40035	2769	0.54
Creek	40037	11884	2.32
Custer	40039	3830	0.75
Delaware	40041	3930	0.77
Dewey	40043	833	0.16
Ellis	40045	715	0.14
Garfield	40047	10120	1.98
Garvin	40049	5017	0.98
Grady	40051	6401	1.25
Grant	40053	668	0.13
Greer	40055	1300	0.25
Harmon	40057	855	0.17
Harper	40059	745	0.15
Haskell	40061	1474	0.29
Hughes	40063	1966	0.38
Jackson	40065	5394	1.05

Jefferson	40067	1143	0.22
Johnston	40069	1775	0.35
Kay	40071	7237	1.41
Kingfisher	40073	2674	0.52
Kiowa	40075	2272	0.44
Latimer	40077	1404	0.27
Le Flore	40079	3730	0.73
Lincoln	40081	4517	0.88
Logan	40083	4574	0.89
Love	40085	1237	0.24
McClain	40087	4467	0.87
McCurtain	40089	3379	0.66
McIntosh	40091	3282	0.64
Major	40093	1339	0.26
Marshall	40095	2271	0.44
Mayes	40097	5819	1.14
Murray	40099	2036	0.40
Muskogee	40101	11395	2.22
Noble	40103	1691	0.33
Nowata	40105	1234	0.24
Okfuskee	40107	1698	0.33
Oklahoma	40109	98887	19.30
Okmulgee	40111	5799	1.13
Osage	40113	6034	1.18
Ottawa	40115	5182	1.01
Pawnee	40117	2780	0.54
Payne	40119	8514	1.66
Pittsburg	40121	7447	1.45
Pontotoc	40123	4975	0.97
Pottawatomie	40125	9709	1.90
Pushmataha	40127	2089	0.41
Roger Mills	40129	577	0.11
Rogers	40131	11181	2.18
Seminole	40133	3206	0.63
Sequoyah	40135	2247	0.44
Stephens	40137	7312	1.43
Texas	40139	1828	0.36
Tillman	40141	1449	0.28
Tulsa	40143	79767	15.57
Wagoner	40145	8155	1.59
Washington	40147	6268	1.22

Washita	40149	2034	0.40
Woods	40151	1285	0.25
Woodward	40153	3151	0.62
Out of state	99998	9682	1.89