# GeoService

# Contents

* 1 [Contents](#GeoService-Contents)
* 2 [Conventions and Guidelines](#GeoService-ConventionsandGuidelines)
  + 2.1 [Column naming](#GeoService-Columnnaming)
* 3 [CDW-H Queries and Functions](#GeoService-CDW-HQueriesandFunctions)
* 4 [GeoService Implementation Process](#GeoService-GeoServiceImplementationProc)
  + 4.1 [Phase 1 – Geocode Patients](#GeoService-Phase1–GeocodePatients)
    - 4.1.1 [Status: Complete](#GeoService-Status:Complete)
  + 4.2 [Phase 2 – Load patient geocode data](#GeoService-Phase2–Loadpatientgeocodedat)
    - 4.2.1 [Status: Complete](#GeoService-Status:Complete.1)
  + 4.3 [Phase 3 – Link Patient geocode data socioeconomic data](#GeoService-Phase3–LinkPatientgeocodedat)
    - 4.3.1 [Status: Complete](#GeoService-Status:Complete.2)
  + 4.4 [Phase 4 – Create ongoing process](#GeoService-Phase4–Createongoingprocess)
    - 4.4.1 [Status: In Progress](#GeoService-Status:InProgress)

# Conventions and Guidelines

**The GEO schema is for the lookup and reference tables, please do not use the GEO schema for researcher data.**

The IDSci geo service tables within the : GEO schema e.g. GEO.PAT\_ADDR\_CHNG\_HX and are located within the SQL Server TRACS\_CUSTOM space

Reference Tables have the REF prefix e.g. GEO.REF\_ZIPCODES or GEO.REF\_UNCH\_DEPARTMENTS

GeoService files and documents located here: **ADAM LET'S DISCUSS WHETHER THESE SHOULD BE INCORPORATED INTO CONFLUENCE IN KEEPING WITH OUR GOAL OF CENTRALIZING DOCUMENTATION. Large files may not be best for Confluence like ACS data files**  
J:\Biomedical Informatics\Carolina Data Warehouse for Health\GeoService

## Column naming

Due to having multiple sources of reference data, some terminology is not consistent in reference files. Appropriate columns shall be called out in each data dictionary for name changes and justification. In general, UNC centric naming conventions override reference sources. An example is the use of CENSUS\_TRACT, this concept is the defined nomenclature for naming a 11 digit string that is use to identify a U.S. Census tract. The US Gazetteer files call this field GEOID, while RUCA files call this FIPS. They shall exist in the source files as originally sourced, but renamed for UNC usages.

# CDW-H Queries and Functions

**GEO\_CODE\_PAT** Expand source

SELECT DISTINCT top 10000

PAT.PAT\_ID

, PAT\_ADDR\_CHNG\_HX.EFF\_START\_DATE

, PAT\_ADDR\_CHNG\_HX.EFF\_END\_DATE

, PAT\_ADDR\_CHNG\_HX.ADDR\_HX\_LINE1

, PAT\_ADDR\_CHNG\_HX.ADDR\_HX\_LINE2

, PAT\_ADDR\_CHNG\_HX.CITY\_HX

, PAT\_ADDR\_CHNG\_HX.ZIP\_HX

, zcstate.NAME as STATE\_HX

, geo.Longitude

, geo.Latitude

, geo.Geocoding\_Precision

, geo.Census\_Tract

from CLARITY.dbo.PATIENT PAT

INNER JOIN CLARITY.dbo.PAT\_ADDR\_CHNG\_HX PAT\_ADDR\_CHNG\_HX ON PAT\_ADDR\_CHNG\_HX.PAT\_ID = PAT.PAT\_ID

INNER JOIN

(

SELECT

PAT\_ID,

MAX(LINE) as LINE

FROM TRACS\_CUSTOM.[dbo].[GEO\_PAT\_ADDR\_CHNG\_HX]

GROUP BY PAT\_ID

)

mostrecentgeo on mostrecentgeo.PAT\_ID = PAT\_ADDR\_CHNG\_HX.PAT\_ID and mostrecentgeo.LINE = PAT\_ADDR\_CHNG\_HX.LINE

INNER JOIN TRACS\_CUSTOM.[dbo].[GEO\_PAT\_ADDR\_CHNG\_HX] geo on geo.PAT\_ID = mostrecentgeo.PAT\_ID and geo.LINE = mostrecentgeo.LINE

LEFT JOIN ZC\_STATE zcstate on zcstate.STATE\_C = PAT\_ADDR\_CHNG\_HX.STATE\_HX\_C

where PAT.PAT\_ID = ''

**GEO\_CODE\_ENC** Expand source

SELECT DISTINCT top 10000

PAT\_ENC.PAT\_ID

, PAT\_ENC.PAT\_ENC\_CSN\_ID

, PAT\_ENC.EFFECTIVE\_DATE\_DT

, PAT\_ADDR\_CHNG\_HX.EFF\_START\_DATE

, PAT\_ADDR\_CHNG\_HX.EFF\_END\_DATE

, PAT\_ADDR\_CHNG\_HX.ADDR\_HX\_LINE1

, PAT\_ADDR\_CHNG\_HX.ADDR\_HX\_LINE2

, PAT\_ADDR\_CHNG\_HX.CITY\_HX

, PAT\_ADDR\_CHNG\_HX.ZIP\_HX

, zcstate.NAME as STATE\_HX

, geo.Longitude

, geo.Latitude

, geo.Geocoding\_Precision

, geo.Census\_Tract

from CLARITY.dbo.PAT\_ENC PAT\_ENC

LEFT JOIN CLARITY.dbo.PAT\_ADDR\_CHNG\_HX PAT\_ADDR\_CHNG\_HX ON PAT\_ADDR\_CHNG\_HX.PAT\_ID = PAT\_ENC.PAT\_ID

AND PAT\_ENC.EFFECTIVE\_DATE\_DT between CASE WHEN PAT\_ADDR\_CHNG\_HX.LINE = 1 THEN '1900-01-01' ELSE PAT\_ADDR\_CHNG\_HX.EFF\_START\_DATE END

AND DATEADD(DAY, -1, COALESCE(PAT\_ADDR\_CHNG\_HX.EFF\_END\_DATE,'2099-01-01'))

LEFT JOIN TRACS\_CUSTOM.[dbo].[GEO\_PAT\_ADDR\_CHNG\_HX] geo on geo.PAT\_ID = PAT\_ADDR\_CHNG\_HX.PAT\_ID and geo.LINE = PAT\_ADDR\_CHNG\_HX.LINE

LEFT JOIN ZC\_STATE zcstate on zcstate.STATE\_C = PAT\_ADDR\_CHNG\_HX.STATE\_HX\_C

where PAT\_ENC.PAT\_ID = ‘’

order by PAT\_ENC.EFFECTIVE\_DATE\_DT

**Patient to GeoService** Expand source

SELECT DISTINCT top 10000

PAT.PAT\_ID

, geocode.Longitude

, geocode.Latitude

, geocode.Geocoding\_Precision

, geodata.\*

from CLARITY.dbo.PATIENT PAT

INNER JOIN CLARITY.dbo.PAT\_ADDR\_CHNG\_HX PAT\_ADDR\_CHNG\_HX ON PAT\_ADDR\_CHNG\_HX.PAT\_ID = PAT.PAT\_ID

INNER JOIN

(

SELECT

PAT\_ID,

MAX(LINE) as LINE

FROM TRACS\_CUSTOM.[dbo].[GEO\_PAT\_ADDR\_CHNG\_HX]

GROUP BY PAT\_ID

)

mostrecentgeo on mostrecentgeo.PAT\_ID = PAT\_ADDR\_CHNG\_HX.PAT\_ID and mostrecentgeo.LINE = PAT\_ADDR\_CHNG\_HX.LINE

INNER JOIN [TRACS\_CUSTOM].[dbo].[GEO\_PAT\_ADDR\_CHNG\_HX] geocode on geocode.PAT\_ID = mostrecentgeo.PAT\_ID and geocode.LINE = mostrecentgeo.LINE

INNER JOIN [TRACS\_CUSTOM].[GEO].[geodata] geodata on geocode.CENSUS\_TRACT = geodata.CENSUS\_TRACT

where PAT.PAT\_ID = '';

**GeoCode Haversine Formula** Expand source

/\*This is used to calculate the distance between a patient X/Y coordinates and another X/Ycoordinates \*/

SELECT \*, 111.045\* DEGREES(ACOS(COS(RADIANS(LATPOINT))

\* COS(RADIANS(LATITUDE))

\* COS(RADIANS(LONGPOINT) - RADIANS(LONGITUDE))

+ SIN(RADIANS(LATPOINT))

\* SIN(RADIANS(LATITUDE)))) / 1.609344 AS distance\_in\_miles

FROM CDWH..RPT\_D\_PTNT

LEFT OUTER JOIN CDWH..GEOCODE ON RPT\_D\_PTNT.PTNT\_DK = CDWH..GEOCODE.PTNT\_DK

JOIN (

SELECT 35.9053 AS latpoint,

-79.0510 AS longpoint

) AS p ON 1=1 --UNC Coords 101 Manning Dr. Chapel Hill, NC 27514

--WHERE distance\_in\_miles <= 50 -- Uncomment to set miles threshold

LIMIT 100000;

# GeoService Implementation Process

The process of geocoding patient location and loading them into the CDW-H is divided into 4 phases:

## ****Phase 1**** – Geocode Patients

#### ****Status:**** Complete

## ****Phase 2**** – Load patient geocode data

#### ****Status:**** Complete

The geocode data has been processed and loaded into the CDW-H production environment. The initial load of geocoded patients included 3,735,440 records with an 87% match success.

## ****Phase 3**** – Link Patient geocode data socioeconomic data

#### ****Status:**** Complete

The Claritas socioeconomic data needs to be loaded into CDW-H (or other database) for the geocode data to be processed against. This will grant researchers access to census tract information for given patients, such as income range and race / ethnicity ratio.

## ****Phase 4**** – Create ongoing process

#### ****Status:**** In Progress

Patient geocode data needs to be updated quarterly. New patients added to the system and address changes need to be captured and processed. Data will be updated over time creating a longitudinal record of geolocation changes.