LAGOS Lake Link: Creation

The purpose of this document is to delineate the processing steps used to join lake identifiers from 5 data products into a single lake identifier table called LAGOS Lake Link (**working name**).

# Definitions in this document

“Lake”: Permanent lake or reservoir.

# What is LAGOS Lake Link?

LAGOS Lake Link is a tabular dataset (a crosswalk table) that can be used to connect one lake-related dataset to another for many common lake datasets. Several lake datasets are in common use on their own or as a base for scientific data products and LAGOS Lake Link is intended to make it easier to combine lake-related data between multiple sources. The table can be searched to find identifiers and location for a particular lake, or it can be used in data join operations to convert identifiers en masse.

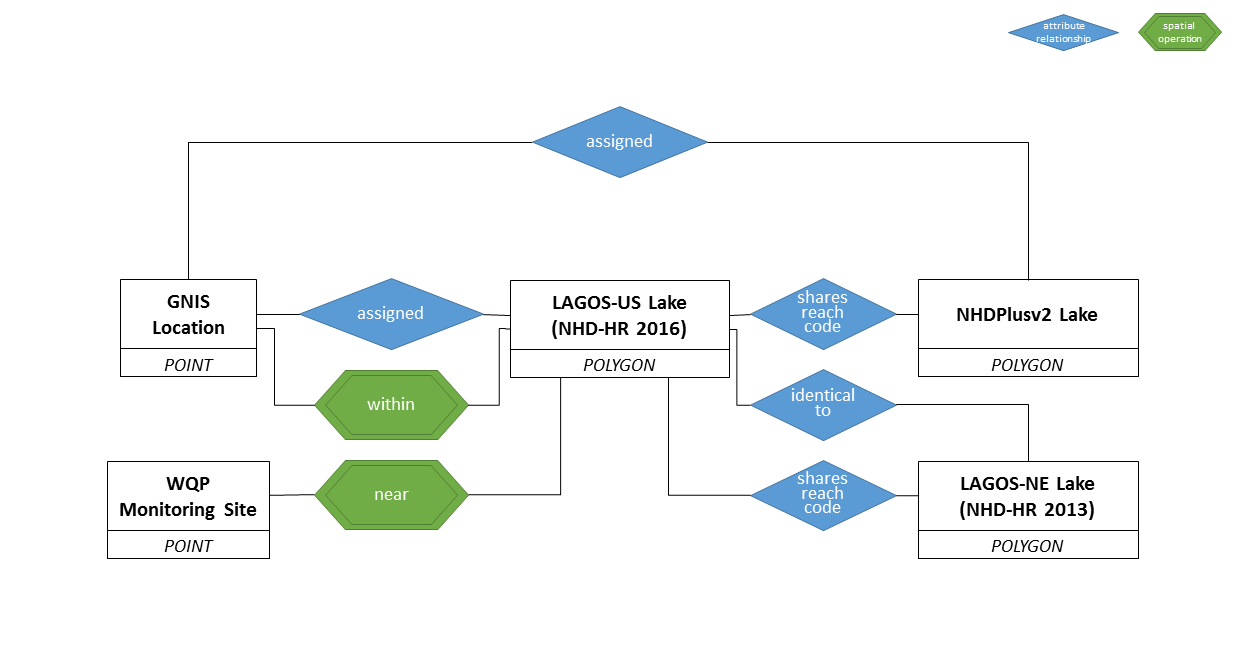
There are two main spatial representations used for lake locations: polygons, used in the various NHD products (National Hydrography Dataset), and points, used in other national datasets. In addition to the spatial representations, there are a variety of identifiers found among the datasets portraying lakes in the United States: state agency IDs, Water Quality Portal MonitoringLocationIdentifier (from systems such as STORET and NWIS), NHDPlusV2 COMID, NHD ReachCode, NHD-HR Permanent\_Identifier, GNIS\_ID. Finally, lakes are sometimes named and can be identified by their name alone if it unique, or their name and geographic context.

There are several obstacles to reliably identifying common lakes between these datasets:

* Inadequate location accuracy
* Inconsistent classification of waterbodies as lakes vs. another type
* Inconsistent delineation of lake extent
* Conflicting names
* Unaccounted change in identifiers within a dataset
* Unsynchronized update patterns between datasets

When I say “inadequate” or “inconsistent” here, I am describing the quality relative to my specific task of connecting datasets. Independently, these datasets have high standards for data quality–but two datasets may use, for example, different definitions of the lake entity, and that is where problems arise for my task.

# Entity-relationship diagram

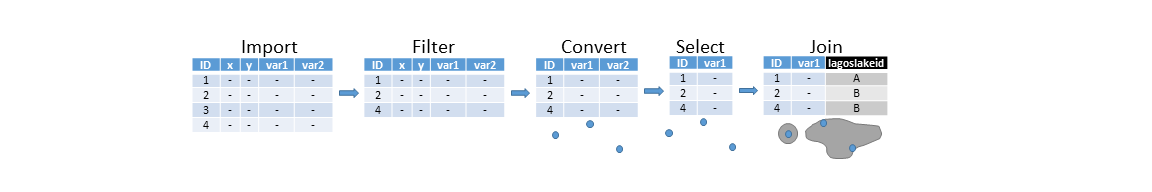


*Entity-relationship diagram*

# Data Processing Pipeline

5 main datasets will be introduced to create the crosswalk: LAGOS (US), GNIS, WQP, NHDPlusV2 (medium-resolution), and LAGOS-NE. For each dataset, the workflow is as follows:

1. **Import.** Modify fields and prepare the data frame for later work.
2. **Filter.** We focus here on lakes and reservoirs in the continental United States. In order to preserve this focus, categories of entities that are less likely to represent one of these features are pruned away during this step, even if those categories sometimes connect to a LAGOS-US lake. The crosswalk only includes relationships if they ultimately connect back to a LAGOS-US lake. For example–even though you can use this table to walk between WQP sites and NHDPlusV2 lakes, the relationship is not comprehensive because lakes not found in LAGOS-US are missing.
3. **Convert** between spatial and non-spatial data formats, as needed. Spatial formats are projected to the USGS Albers Conic Equal Area projection for consistency.
4. **Select** only the necessary columns for the upcoming work.
5. **Join LAGOS-US** to each other dataset, individually. This step may include spatial joins, joins on common identifiers. After joining, identifying fields may be flattened into a single concatenated field or coalesced.
6. **Select again** again as needed before the final join.



*Processing Pipeline*

Finally, the results of the multiple joins in step 4 will be merged into a single crosswalk table with only the necessary fields remaining.

# LAGOS-US

The LAGOS-US lake population is the focal dataset of the LAGOS Lake Link crosswalk (top center, Figure ). We will connect all other datasets to this one in the crosswalk. Lakes will only This dataset is a subset of the NHD High Resolution NHDWaterbody layers (see “Filter” section for details) representing permanent, non-artificial LakePond and Reservoir waterbodies. Currently, LAGOS-US is under development and while the lake polygon dataset is completed, much of the database remains to be built in 2018 and 2019.

The primary identifier for lakes in this dataset is **lagoslakeid**. A secondary unique identifier is named **nhdid** in the “locus” table, which corresponds to the **Permanent\_Identifier** in the NHD and in the GIS layer imported here.

## Import

We’ll import the GIS polygon layer prepared for LAGOS-US. Glimpse the polygon layer:

## Observations: 479,950  
## Variables: 24  
## $ Permanent\_Identifier <chr> "120020498", "107075016", "107071410", "120020448", "120020445", "...  
## $ FDate <dttm> 2004-03-16 13:37:00, 2003-10-09 00:00:00, 2013-04-25 00:00:00, 20...  
## $ Resolution <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...  
## $ GNIS\_ID <int> 649289, 653169, NA, 1776969, 662208, 1579742, NA, NA, NA, 656308, ...  
## $ GNIS\_Name <chr> "Perch Lake", "Thomson Reservoir", NA, "Eveleth Taconite Tailings ...  
## $ AreaSqKm <dbl> 2.74500000, 1.38300000, 1.25090582, 2.16000000, 1.35800000, 1.1440...  
## $ Elevation <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ ReachCode <chr> "04010201003330", "04010201003334", "04010201006300", "04010201006...  
## $ FType <int> 390, 390, 390, 390, 390, 390, 390, 390, 390, 390, 390, 390, 390, 3...  
## $ FCode <int> 39004, 39004, 39004, 39004, 39004, 39004, 39004, 39011, 39004, 390...  
## $ VisibilityFilter <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ VertexCount <int> 541, 550, 126, 432, 969, 664, 71, 46, 467, 488, 4581, 1050, 204, 2...  
## $ HU4 <chr> "0401", "0401", "0401", "0401", "0401", "0401", "0401", "0401", "0...  
## $ HU6 <chr> "040102", "040102", "040102", "040102", "040102", "040103", "04010...  
## $ HU8 <chr> "04010201", "04010201", "04010201", "04010201", "04010201", "04010...  
## $ AreaHa <dbl> 274.537420, 138.341755, 125.090582, 216.040795, 135.835390, 114.43...  
## $ Shape\_Length <dbl> 12649.8933, 14119.9959, 8505.3657, 8422.8245, 9104.7865, 11131.817...  
## $ Shape\_Area <dbl> 2745374.20, 1383417.55, 1250905.82, 2160407.95, 1358353.90, 114437...  
## $ HU12\_1 <chr> "040102011201", "040102011505", "040102010602", "040102010702", "0...  
## $ STATE <chr> "MN", "MN", "MN", "MN", "MN", "WI", "MN", "MN", "MN", "MN", "MN", ...  
## $ lagoslakeid <int> 31198, 122, 38509, 105038, 23963, 4674, 118310, 256600, 47947, 270...  
## $ NAMELSAD <chr> "Carlton County", "Carlton County", "St. Louis County", "St. Louis...  
## $ FIPS <chr> "27017", "27017", "27137", "27137", "27137", "55007", "27137", "27...  
## $ Shape <sf\_geometry [m]> MULTIPOLYGON (((254295.3 26..., MULTIPOLYGON (((277079...

## Filter

The LAGOS-US dataset has already been filtered from the source dataset of all NHDWaterbody features to generate the final target lake population. The code is not included but the processing rules are summarized below.

Criteria for selecting NHD lake polygons for inclusion in LAGOS-US were as follows. Included lakes must:

* Be derived from snapshots of the NHD staged by subregion taken 2016-12-15, 2016-12-16, 2017-01-03 (HUC4 1019), 2017-01-03 (HUC4 0309, 1004, 1804).
* Not be an exact duplicate of another NHD lake polygon.
* If a duplicate of another lake except for FDate and Shape, then the lake retained must have the most recent FDate (feature edit date).
* Intersect the LAGOS-US contiguous U.S. layer (48 states plus D.C., TIGER/Line data).
* Be represented as simple features–we densified several features with only 2 vertices using 10 m as the maximum deviation for each vertex in order to “repair” their representation to match the majority and the standard.
* Not be Great Lakes or the portion of Lake St. Clair (MI/ON) that is in Canada.
* Have an AreaSqKm > 0.009 (AreqSqKm field included in original NHDWaterbody).
* Have a polygon area greater than or equal to 1 hectare (calculated during LAGOS-US processing and measured in the Albers USGS Conic projection). The calculated area occasionally does not match the area in square kilometers provided by NHD.
* Be assigned one of the following Feature Codes representing permanent water bodies and non-artificial water bodies:
  + 39000,39004,39009,39010,39011,39012 (lakes)
  + 43600,43613,43615,43617,43618,43619,43621 (reservoirs)
* One lake in Mexico was removed (Permanent\_Identifier = ‘e05e57b5-d29f-4e1e-8369-73e55f8be9df’) after independent verification of the lake processing workflow revealed a single discrepancy due to slight variations in order of when data projection was performed.

The **number of LAGOS-US lakes** after filtering is **479950**.

## Convert and Select

We will select several columns and create two alternate representations of the LAGOS-US dataset: centroids and no geometry.

lagos <- lagos\_sf %>%  
 select(lagoslakeid,   
 Permanent\_Identifier,   
 GNIS\_ID,   
 GNIS\_Name,   
 ReachCode,   
 STATE,   
 AreaSqKm,   
 FDate,  
 LAGOS\_CountyName = NAMELSAD,  
 LAGOS\_CountyFIPS = FIPS)  
  
lagos\_centroid <- lagos\_sf %>%  
 st\_centroid() %>%  
 st\_transform(NAD83) %>%  
 mutate(LAGOS\_LatitudeNAD83 = st\_coordinates(.)[,'Y'],   
 LAGOS\_LongitudeNAD83 = st\_coordinates(.)[,'X']) %>%  
 select(lagoslakeid, LAGOS\_LatitudeNAD83, LAGOS\_LongitudeNAD83)  
  
lagos\_df <- lagos %>% st\_set\_geometry(NULL)

# GNIS (Geographic Names Information System)

The GNIS dataset contains place names for a variety of geographic features, including lakes, in the United States. Locations are represented as latitude/longitude pairs. Most NHD lakes already have a GNIS name assigned, but in some cases where lakes had multiple names, they were not assigned a name in the NHD. To populate the lake name as fully as possible, we searched the GNIS dataset to both add missing names where available and to document lakes with multiple names, which we then retain in LAGOS-US.

## Import

Glimpse the original GNIS table.

## Observations: 2,276,506  
## Variables: 20  
## $ FEATURE\_ID <int> 0, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412...  
## $ FEATURE\_NAME <chr> "Bruceville Cemetery", "Agua Sal Creek", "Agua Sal Wash", "Aguaje Draw"...  
## $ FEATURE\_CLASS <chr> "Cemetery", "Stream", "Valley", "Valley", "Park", "Stream", "Stream", "...  
## $ STATE\_ALPHA <chr> "CO", "AZ", "AZ", "AZ", "AZ", "AZ", "AZ", "AZ", "AZ", "AZ", "AZ", "AZ",...  
## $ STATE\_NUMERIC <chr> "08", "04", "04", "04", "04", "04", "04", "04", "04", "04", "04", "04",...  
## $ COUNTY\_NAME <chr> "Boulder", "Apache", "Apache", "Apache", "Maricopa", "Graham", "Apache"...  
## $ COUNTY\_NUMERIC <chr> "013", "001", "001", "001", "013", "009", "001", "019", "017", "005", "...  
## $ PRIMARY\_LAT\_DMS <chr> "400058N", "362740N", "363246N", "343417N", "331455N", "322815N", "3552...  
## $ PRIM\_LONG\_DMS <chr> "1051226W", "1092842W", "1093103W", "1091313W", "1124625W", "1095610W",...  
## $ PRIM\_LAT\_DEC <dbl> 40.01610, 36.46111, 36.54611, 34.57143, 33.24865, 32.47090, 35.87501, 3...  
## $ PRIM\_LONG\_DEC <dbl> -105.2072, -109.4784, -109.5176, -109.2204, -112.7735, -109.9362, -109....  
## $ SOURCE\_LAT\_DMS <chr> NA, "362053N", "362740N", "344308N", NA, "323048N", "354903N", "315820N...  
## $ SOURCE\_LONG\_DMS <chr> NA, "1090915W", "1092842W", "1085826W", NA, "1095233W", "1093001W", "11...  
## $ SOURCE\_LAT\_DEC <dbl> NA, 36.34806, 36.46111, 34.71880, NA, 32.51340, 35.81750, 31.97230, 34....  
## $ SOURCE\_LONG\_DEC <dbl> NA, -109.1543, -109.4784, -108.9739, NA, -109.8759, -109.5003, -111.391...  
## $ ELEV\_IN\_M <int> 1598, 1645, 1597, 1750, 231, 1339, 1799, 591, 1494, 1435, 1506, 81, 144...  
## $ ELEV\_IN\_FT <int> 5243, 5397, 5239, 5741, 758, 4393, 5902, 1939, 4902, 4708, 4941, 266, 4...  
## $ MAP\_NAME <chr> "Niwot", "Fire Dance Mesa", "Little Round Rock", "Kearn Lake", "Spring ...  
## $ DATE\_CREATED <chr> "05/04/2015", "02/08/1980", "02/08/1980", "02/08/1980", "02/08/1980", "...  
## $ DATE\_EDITED <chr> "10/22/2015", NA, NA, "01/14/2008", NA, NA, NA, NA, NA, NA, NA, NA, NA,...

## Filter

Included GNIS sites must:

* Be within one of the 48 contiguous states or D.C (FIPS code match)
* Be within the rough bounding box of the U.S. (eliminates a few spurious locations)
* Be assigned the Lake or Reservoir Feature Class. These feature classes aggregate a variety of feature names such as Lake, Lakes, Reservoir, Pond, Tank, Slough, Millpond, etc. *A prior analysis showed that nearly all GNIS names already assigned in the NHD correspond to one of these two classes, with minor use of Swamp, Flat, and rare use of a smattering of other feature classes. We choose to focus the crosswalk on the most relevant locations rather than conserve all possible links.*
* Not be assigned a “historical” site name. Such names are spatially coincident with the current name.

The **number of GNIS sites** for U.S. lakes is **138055**.

## Convert

According to the [metadata](https://geonames.usgs.gov/domestic/faqs.htm), the coordinates use the NAD83 datum. Convert the XY data to points.

## Select

Select the FEATURE\_ID, FEATURE\_NAME, and FEATURE\_CLASS.

gnis <- gnis\_sf %>%  
 select(FEATURE\_ID, FEATURE\_NAME, FEATURE\_CLASS)

## Join LAGOS-US

First, join LAGOS-US to GNIS based on the shared GNIS\_ID.

GNIS Names contained by the polygon are assigned to the lake. A new column, “LAGOS Lake Name” is created and multiple names for a lake are separated by semi-colons. This will eliminate the 1:many relationship created in the join.

## Select again

Select only the desired names and identifiers.

lagos\_gnis <- lagos\_gnis\_3 %>%  
 st\_set\_geometry(NULL) %>%  
 select(lagoslakeid, Permanent\_Identifier, GNIS\_ID, GNIS\_Name, ReachCode, STATE, LAGOS\_LakeName, AreaSqKm)

# WQP (Water Quality Portal–STORET, NWIS, and more)

The WQP API allows queries of all water quality monitoring sites that are submitted to major national databases. Sites are described with identifiers, coordinates, and names, but are not definitively linked to any polygon representation of lakes to our knowledge. The goal here is to match the WQP identifiers with lakes in LAGOS-US.

## Import

Glimpse the original WQP table.

## Observations: 2,300,409  
## Variables: 36  
## $ OrganizationIdentifier <chr> "ARS", "ARS", "ARS", "ARS", "ARS", "ARS...  
## $ OrganizationFormalName <chr> "Agricultural Research Service", "Agric...  
## $ MonitoringLocationIdentifier <chr> "ARS-IASF-IABC264", "ARS-IASF-IABC274",...  
## $ MonitoringLocationName <chr> "IABC264", "IABC274", "IABC350", "IASF2...  
## $ MonitoringLocationTypeName <chr> "Land Runoff", "Land Runoff", "River/St...  
## $ MonitoringLocationDescriptionText <chr> "A drainage ditch monitoring site on 19...  
## $ HUCEightDigitCode <chr> "07080207", "07080207", "07080207", "07...  
## $ DrainageAreaMeasure.MeasureValue <dbl> NA, NA, 18210.0, NA, NA, NA, NA, NA, NA...  
## $ DrainageAreaMeasure.MeasureUnitCode <chr> NA, NA, "ha", NA, NA, NA, NA, NA, NA, N...  
## $ ContributingDrainageAreaMeasure.MeasureValue <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ ContributingDrainageAreaMeasure.MeasureUnitCode <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ LatitudeMeasure <dbl> 42.41918, 42.40710, 42.33911, 42.54317,...  
## $ LongitudeMeasure <dbl> -93.18051, -93.20679, -93.12500, -93.47...  
## $ SourceMapScaleNumeric <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ HorizontalAccuracyMeasure.MeasureValue <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ HorizontalAccuracyMeasure.MeasureUnitCode <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ HorizontalCollectionMethodName <chr> "GPS Code (Pseudo Range) Differential",...  
## $ HorizontalCoordinateReferenceSystemDatumName <chr> "NAD83", "NAD83", "NAD83", "NAD83", "NA...  
## $ VerticalMeasure.MeasureValue <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ VerticalMeasure.MeasureUnitCode <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ VerticalAccuracyMeasure.MeasureValue <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ VerticalAccuracyMeasure.MeasureUnitCode <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ VerticalCollectionMethodName <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ VerticalCoordinateReferenceSystemDatumName <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ CountryCode <chr> "US", "US", "US", "US", "US", "US", "US...  
## $ StateCode <chr> "19", "19", "19", "19", "19", "19", "19...  
## $ CountyCode <chr> "083", "083", "083", "083", "079", "083...  
## $ AquiferName <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ FormationTypeText <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ AquiferTypeName <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ ConstructionDateText <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ WellDepthMeasure.MeasureValue <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ WellDepthMeasure.MeasureUnitCode <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ WellHoleDepthMeasure.MeasureValue <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ WellHoleDepthMeasure.MeasureUnitCode <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ ProviderName <chr> "STEWARDS", "STEWARDS", "STEWARDS", "ST...

## Filter

Included WQP sites must:

* Be within the continental United States (rough cut to eliminate distant territories)
* Be assigned one of the following Monitoring Location Types: “Lake”, “Reservoir”, “Riverine Impoundment”, “Lake, Reservoir, Impoundment”. *These categories were selected in an analysis that assessed how often the sites of each type were located within an NHD polygon and whether they often had Secchi depth measurements, a characteristically limnological sampling parameter. Other location types did sometimes represent desirable samples, but we choose to focus the crosswalk on the most relevant locations rather than conserve all possible links. Stream, wetlands, estuaries and Great Lakes were categorically excluded despite meeting the above criteria in some part.*

The **number of WQP sites** meeting these criteria is **131005**.

## Convert

There is a column in WQP data to indicate the coordinate reference system (CRS). However, not all sites had a value in this column. Prior analysis showed that the most commonly indicated column was NAD83, and so for all sites with an unknown or rare CRS (distant U.S. territories), the NAD83 datum was imputed instead.

For sites with one of the following other datums, the actual datum specification is preserved: NAD27, WGS84. In order to manage the projections before transforming all the data to a common projection, Albers Equal Area Conic (USGS), we split up the dataset, projected the sites in each datum, converted all rows to Albers USGS, and then reunited all rows at the end.

## Select

wqp <- wqp\_sf %>%  
 select(OrganizationIdentifier, MonitoringLocationIdentifier, MonitoringLocationName, ProviderName, MonitoringLocationTypeName)

## Join LAGOS-US

1. WQP sites within a lake polygon are linked to that polygon.

Examination of shared words in the names: All lakes where two names are available but don’t share words. Some of these lakes have names following some kind of technical form, but others seems to suggest mis-referenced lakes even though the point for the site falls inside the lake. Some examples (zoom in to see NHD detail appear):

1. WQP sites can be joined to a lake polygon if they are within 10 meters of it.
2. Other sites can be joined if they are within 100 m of a lake polygon, not within an NHDArea StreamRiver feature, and share name words other than “lake”, “reservoir”, etc.

Join them all up into one and enforce the rules.

How many joins were made for each method?

* **No search distance:** 91958
* **10 m search distance:** 5139
* **100 m search distance (with name condition):** 5267

Some lakes have a very high number of sampling sites. This map shows some examples of lakes with over 50 sampling sites stored.

## Select again

Retain only columns relevant to crosswalk.

lagos\_wqp <- lagos\_wqp\_all %>%  
 st\_set\_geometry(NULL) %>%  
 select(OrganizationIdentifier, MonitoringLocationIdentifier, MonitoringLocationName, ProviderName, Linked\_lagoslakeid)

# NHDPlusV2 (value-added National Hydrography Dataset and basis for National Lakes Assessment, LakeCat)

The medium-resolution NHD contains a popular medium-resolution representation of lakes and their hydrographic context. Because the high-resolution NHD data were used to generate LAGOS-US, we must find a connection between the datasets. This connection is the reach code, a 14-digit identifier used to identify each reach as part of a linear referencing system. ReachCodes must sometimes be modified (split, joined, re-assigned, deleted). The NHDReachCrossReference table tracks these changes and shows the new and old ReachCode.

## Import

The version of the NHDReachCrossReference table that we use has already been filtered to include only the reach codes for lakes and reservoirs in the NHD HR (e.g. streams, wetlands, etc. were removed) to reduce the table size. You could use the full original table with no change to the results.

Glimpse the NHD table.

## Observations: 448,512  
## Variables: 22  
## $ COMID <int> 806161, 806167, 806169, 806175, 806179, 806181, 806183, 806189, 806197, 80...  
## $ FDATE <dttm> 1999-05-27, 1999-05-27, 1999-05-27, 1999-05-27, 1999-05-27, 1999-05-27, 1...  
## $ RESOLUTION <chr> "Medium", "Medium", "Medium", "Medium", "Medium", "Medium", "Medium", "Med...  
## $ GNIS\_ID <dbl> NA, NA, 564483, 563828, 575611, NA, NA, NA, NA, 572518, NA, NA, 563822, NA...  
## $ GNIS\_NAME <chr> " ", " ", "Cranberry Pond", "Chandler Lake", "Smith Brook Pond", " ", " ",...  
## $ AREASQKM <dbl> 0.013, 0.055, 0.176, 1.572, 0.069, 0.162, 0.136, 0.187, 0.049, 0.018, 0.01...  
## $ ELEVATION <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ REACHCODE <chr> "01010004001814", "01010004001819", "01010004001817", "01010004001854", "0...  
## $ FTYPE <chr> "LakePond", "LakePond", "LakePond", "LakePond", "LakePond", "LakePond", "L...  
## $ FCODE <int> 39004, 39004, 39004, 39004, 39004, 39004, 39004, 39004, 39004, 39004, 3900...  
## $ Shape\_Length <dbl> 0.005468933, 0.021807035, 0.021755760, 0.087375064, 0.017330298, 0.0530475...  
## $ Shape\_Area <dbl> 1.513469e-06, 6.416618e-06, 2.057280e-05, 1.841445e-04, 8.053727e-06, 1.89...  
## $ ONOFFNET <int> 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,...  
## $ PurpCode <chr> " ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ",...  
## $ PurpDesc <chr> " ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ", " ",...  
## $ MeanDepth <dbl> 0.09450471, 0.09872492, 0.28167548, NA, 0.56829879, 0.13899818, 0.55463734...  
## $ LakeVolume <dbl> 1220.5177, 5407.7469, 49470.2953, NA, 39081.4758, 22468.7585, 75691.3313, ...  
## $ MaxDepth <dbl> 0.23044897, 0.23288202, 0.63270356, NA, 1.09330535, 0.27875821, 1.26695295...  
## $ MeanDUsed <dbl> 0.09450471, 0.09872492, 0.28167548, NA, 0.56829879, 0.13899818, 0.55463734...  
## $ MeanDCode <chr> "4", "4", "4", NA, "4", "4", "4", "4", "4", "4", "4", NA, "4", "4", "4", "...  
## $ LakeArea <dbl> 12914.89, 54775.91, 175628.68, NA, 68769.23, 161647.88, 136469.96, 187427....  
## $ Shape <sf\_geometry [degree]> MULTIPOLYGON (((-68.31665 4..., MULTIPOLYGON (((-68.3228 ...

Glimpse the NHDReachCrossReference table.

## Observations: 5,856,425  
## Variables: 4  
## $ OldReachCode <fct> 01010003000640, 01010004001528, 01010004001679, 01010004001536, 0101000400...  
## $ OldReachDate <dttm> 1997-05-21, 1997-05-21, 1997-05-21, 1997-05-21, 1997-05-21, 1997-05-21, 1...  
## $ NewReachCode <fct> 01010003000737, 01010004001871, 01010004001873, 01010004001874, 0101000400...  
## $ NewReachDate <dttm> 1998-07-04, 1998-07-04, 1998-07-04, 1998-07-04, 1998-07-04, 1998-07-04, 1...

## Filter

We can’t use the join method to connect to any NHDPlusV2 lakes that don’t have REACHCODE populated. We will revisit these lakes along with others that won’t connect later.

## Convert

These data are already spatial polygons and we want a dataset with no geometry instead.

## Select

nhd\_plus <- nhd\_plus\_df %>%  
 select(COMID, FDATE, GNIS\_ID, GNIS\_NAME, REACHCODE, AREASQKM)

## Join LAGOS-US

This join creates many:many relationships. A scant few ReachCodes in the LAGOS-US population are associated with two Permanent\_Identifiers before the join. Most reach codes don’t have an entry in the cross reference table, suggesting that lakes still holding their original reach code should have the same reach code in any version of the NHD.

**346721** reach codes have only a new reach code in NHDReachCrossReference and the OldReachCode is missing. In other words, about **72.24%** of LAGOS-US lakes have never changed reach codes. About a tenth of the lakes have no entry in NHDReachCrossReference, suggesting this table is usually but not always populated when new reach codes are assigned.

Numerous lakes aren’t able to be connected between the NHDPlusV2 and the NHD HR. The HR contains more small lakes, so let’s look for a minute only at lakes over 10 hectares. Here are a few examples of lakes where the old reach codes are unavailable or not the target reach code.

## OldReachCode OldReachDate NewReachCode NewReachDate  
## 1 01090001001553 1997-05-21 01090001001634 1998-07-04  
## 2 <NA> 1999-11-30 01090001001634 2015-02-27  
## 3 <NA> <NA> 01090003008075 2004-07-28  
## 4 <NA> 1997-05-08 02030104005251 2007-10-23

For those lakes that didn’t link between NHD versions, try linking them through GNIS IDs.

It works. **5316** lakes that couldn’t be linked with NHDReachCrossReference can be linked with the GNIS\_ID value.

## Select again

lagos\_nhd\_not\_missing <- lagos\_nhd\_all %>%  
 filter(!is.na(COMID))  
  
lagos\_nhd <- lagos\_df %>%  
 left\_join(lagos\_nhd\_not\_missing, by = "lagoslakeid") %>% # removes OldReachCodes that don't link but keeps all lakes  
 left\_join(nhd\_gnis\_nhd, by = "lagoslakeid") %>%  
 mutate(NHDPlusv2\_COMID = coalesce(COMID.x, COMID.y)) %>%  
 mutate(NHDPlusv2\_ReachCode = coalesce(OldReachCode, REACHCODE)) %>%  
 mutate(NHDPlusv2\_GNIS\_Name = coalesce(GNIS\_NAME.x, GNIS\_NAME.y)) %>% #all-caps is medium res  
 mutate(NHDPlusv2\_AreaSqKm = coalesce(AREASQKM.x, AREASQKM.y)) %>%  
 select(lagoslakeid, NHDPlusv2\_COMID, NHDPlusv2\_ReachCode, NHDPlusv2\_GNIS\_Name, NHDPlusv2\_AreaSqKm)

# LAGOS-NE

LAGOS-NE was the first release from the LAGOS project describing lakes in 17 states in the northeastern states. It is similar in most ways to LAGOS-US, but the NHD snapshots were taken several years ago. This dataset can be downloaded and unzipped from [<https://portal.edirepository.org/nis/mapbrowse?packageid=edi.98.1>].

When the LAGOS-US lagoslakeid values were generated, they were joined on Permanent\_Identifier to the LAGOS-NE, so most of the lagoslakeid values should not have changed. Nonetheless, some could not be connected this way.

## Import

Glimpse the original LAGOS-NE dataset.

## Observations: 141,265  
## Variables: 35  
## $ ComID <int> 145061933, 145050791, 145050750, 145050563, 142825415, 142825576, ...  
## $ Permanent\_Identifier <fct> 145061933, 145050791, 145050750, 145050563, 142825415, 142825576, ...  
## $ FDate <dttm> 2006-01-11 16:55:46, 2006-01-11 15:27:23, 2006-01-11 15:26:37, 20...  
## $ GNIS\_ID <dbl> NA, 4781, 4907, NA, NA, 5045, NA, 3906, NA, NA, NA, NA, 4214, NA, ...  
## $ GNIS\_Name <fct> NA, Lower Hudson Pond, Mink Marsh Pond, NA, NA, Mud Pond, NA, Deep...  
## $ AreaSqKm <dbl> 0.039, 0.355, 0.073, 0.017, 0.021, 0.066, 0.018, 0.025, 0.014, 0.0...  
## $ Elevation <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ ReachCode <fct> 01010005000620, 01010004001847, 01010004001838, 01010004005842, 01...  
## $ FType <int> 390, 390, 390, 390, 390, 390, 390, 390, 390, 390, 390, 390, 390, 3...  
## $ FCode <int> 39004, 39004, 39004, 39004, 39004, 39004, 39004, 39004, 39004, 390...  
## $ Lake\_Area\_Ha <dbl> 3.888402, 35.450899, 7.317129, 1.684377, 2.088132, 6.614470, 1.817...  
## $ LakeConnectivity <fct> DR\_Stream, DR\_LakeStream, DR\_Stream, Headwater, Headwater, DR\_Stre...  
## $ HU4\_ZoneID <fct> HU4\_1, HU4\_1, HU4\_1, HU4\_1, HU4\_1, HU4\_1, HU4\_1, HU4\_1, HU4\_1, HU4...  
## $ HU4\_Name <fct> St. John, St. John, St. John, St. John, St. John, St. John, St. Jo...  
## $ HU6\_ZoneID <fct> HU6\_1, HU6\_1, HU6\_1, HU6\_1, HU6\_1, HU6\_1, HU6\_1, HU6\_1, HU6\_1, HU6...  
## $ HU6\_Name <fct> St. John, St. John, St. John, St. John, St. John, St. John, St. Jo...  
## $ HU8\_ZoneID <fct> HU8\_5, HU8\_4, HU8\_4, HU8\_4, HU8\_601, HU8\_601, HU8\_601, HU8\_5, HU8\_...  
## $ HU8\_Name <fct> Meduxnekeag, Aroostook, Aroostook, Aroostook, St. Francis River-Sa...  
## $ HU12\_ZoneID <fct> HU12\_17971, HU12\_17917, HU12\_17915, HU12\_17946, HU12\_17762, HU12\_1...  
## $ HU12\_Name <fct> South Branch Meduxnekeag River, Mooseleuk Lake, Bartlett Brook-Spr...  
## $ EDU\_ZoneID <fct> EDU\_79, EDU\_79, EDU\_79, EDU\_79, EDU\_79, EDU\_79, EDU\_79, EDU\_79, ED...  
## $ EDU\_Name <fct> Upper St. John - Aroostook, Upper St. John - Aroostook, Upper St. ...  
## $ COUNTY\_ZoneID <fct> County\_303, County\_312, County\_312, County\_303, County\_303, County...  
## $ COUNTY\_Name <fct> Aroostook County, Piscataquis County, Piscataquis County, Aroostoo...  
## $ State\_Name <fct> Maine, Maine, Maine, Maine, Maine, Maine, Maine, Maine, Maine, Mai...  
## $ STATE\_ZoneID <fct> State\_1, State\_1, State\_1, State\_1, State\_1, State\_1, State\_1, Sta...  
## $ Shape\_Length <dbl> 997.0809, 3626.5048, 2472.0295, 1602.0563, 1056.0734, 1325.3148, 8...  
## $ Shape\_Area <dbl> 38884.02, 354508.99, 73171.29, 16843.77, 20881.32, 66144.70, 18171...  
## $ HU4 <fct> 0101, 0101, 0101, 0101, 0101, 0101, 0101, 0101, 0101, 0101, 0101, ...  
## $ HU6 <fct> 010100, 010100, 010100, 010100, 010100, 010100, 010100, 010100, 01...  
## $ HU12 <fct> 010100050102, 010100040303, 010100040301, 010100040904, 0101000804...  
## $ HU8 <fct> 01010005, 01010004, 01010004, 01010004, 01010008, 01010008, 010100...  
## $ STATE <fct> ME, ME, ME, ME, ME, ME, ME, ME, ME, ME, ME, ME, ME, ME, ME, ME, ME...  
## $ lagoslakeid <int> 29903, 79876, 52999, 97563, 89329, 125499, 132198, 131140, 133209,...  
## $ Shape <sf\_geometry [m]> MULTIPOLYGON (((2148592 288..., MULTIPOLYGON (((204595...

## Filter

No need to filter. These lakes had essentially the same criteria for inclusion as LAGO-US, above, except that they were confined to a 17-state region in the northeast U.S.

## Convert and select

lagosNE <- lagosNE\_orig %>%  
 select(LAGOSNE\_lagoslakeid = lagoslakeid, LAGOSNE\_PermanentIdentifier = Permanent\_Identifier, LAGOSNE\_ReachCode = ReachCode, GNIS\_ID, GNIS\_Name) %>%  
 st\_set\_geometry(NULL)

## Join LAGOS-US

Unlike with NHDPlusV2, these two datasets shared the identifier Permanent\_Identifier and we will try to join lakes on that column first. The subsequent two joins parallel the joins with NHDPlusV2, above. We will reuse the lagos\_nhdx table.

After trying the code below (currently commented out), it seems upon further observation that the GNIS\_ID values in LAGOS-NE have no useful relation to the latest GNIS\_ID values.

## Select again

lagosUS\_lagosNE <- lagos\_lagosNE\_all %>%  
 filter(!is.na(LAGOSNE\_lagoslakeid)) %>%  
 arrange(lagoslakeid)

# NLA (US EPA National Lakes Assessment)

The lakes in the NLA are identified with their NHDPlus COMID and ReachCode. It *should* be simple to add them to the crosswalk by doing a 1:1 join with the NHDPlusV2 COMID.

## Import

Import the two NLA files. It seems that the 2012 file already manages the link between the 2012 sites back to the 2007 sites.

## Observations: 2,764  
## Variables: 111  
## $ SITE\_ID <chr> "NLA12\_AL-101", "NLA12\_AL-101", "NLA12\_AL-102", "NLA12\_AL-103", "NLA12...  
## $ VISIT\_NO <int> 1, 2, 1, 1, 1, 1, 2, NA, NA, NA, NA, NA, NA, 1, 1, 1, 1, 1, 1, 2, 1, 1...  
## $ UID <int> 6300, 7374, 7228, 7292, 7368, 6267, 7394, NA, NA, NA, NA, NA, NA, 7375...  
## $ DATE\_COL <chr> "5/24/2012", "7/21/2012", "7/17/2012", "7/19/2012", "7/23/2012", "5/23...  
## $ AGGR\_ECO3\_2015 <chr> "EHIGH", "EHIGH", "EHIGH", "EHIGH", "PLNLOW", "PLNLOW", "PLNLOW", "PLN...  
## $ AGGR\_ECO9\_2015 <chr> "SAP", "SAP", "SAP", "SAP", "CPL", "CPL", "CPL", "CPL", "CPL", "SAP", ...  
## $ AREA\_HA <dbl> 1301.871970, 1301.871970, 72.950420, 10363.258480, 4731.474336, 32.561...  
## $ AREACLS <chr> "F:500+", "F:500+", "D:50-100", "F:500+", "F:500+", "C:10-50", "C:10-5...  
## $ BORD\_LAKE <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"...  
## $ CAT\_UNIQUE <chr> "NLA07RVT", "NLA07RVT", "NLA07RVT", "NLA07RVT", "NLA07RVT", "NLA12NEW"...  
## $ CH0712\_CAT <chr> "NLA07\_RVT12", "NLA07\_RVT12", "NLA07\_RVT12", "NLA07\_RVT12", "NLA07\_RVT...  
## $ CH0712\_USE <chr> "Include", "Include", "Include", "Include", "Include", "Include", "Inc...  
## $ CH0712\_WGT <dbl> 20.288885, 20.288885, 46.617902, 51.331899, 101.491046, 83.376963, 83....  
## $ CNTYNAME <chr> "TUSCALOOSA", "TUSCALOOSA", "CULLMAN", "TALLADEGA", "SUMTER", "WASHING...  
## $ COMID2007 <int> 21656744, 21656744, 18587568, 22273142, 21677972, 21639448, 21639448, ...  
## $ COMID2012 <int> 166759236, 166759236, 18587568, 166759059, 21677972, 21639448, 2163944...  
## $ COMIDS2007 <chr> "21656744", "21656744", "18587568", "22273142", "21677972", "21639448"...  
## $ DES\_FTYPE <chr> "Lake/Pond: Hydrographic Category = Perennial; Stage = Average Water E...  
## $ DSGN12 <chr> "Included", "Included", "Included", "Included", "Included", "Included"...  
## $ ELEVATION <dbl> 57.03, 57.03, 247.04, 141.76, 22.28, 50.83, 50.83, 96.12, 77.74, 181.4...  
## $ EPA\_REG <chr> "Region\_4", "Region\_4", "Region\_4", "Region\_4", "Region\_4", "Region\_4"...  
## $ EVAL\_NAME <chr> "Holt Lock and Dam", "Holt Lock and Dam", "George Lake", "Logan Martin...  
## $ EVALSTAT <chr> "TargetSampleable", "TargetSampleable", "TargetSampleable", "TargetSam...  
## $ FEOW\_ID <int> 153, 153, 153, 153, 153, 153, 153, 154, 153, 153, 154, 152, 155, 153, ...  
## $ FS\_EW <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ FW\_ECO3 <chr> "EHIGH", "EHIGH", "EHIGH", "EHIGH", "PLNLOW", "PLNLOW", "PLNLOW", "PLN...  
## $ FW\_ECO9 <chr> "SAP", "SAP", "SAP", "SAP", "CPL", "CPL", "CPL", "CPL", "CPL", "SAP", ...  
## $ GNIS\_ID <int> NA, NA, 131985, NA, NA, 117983, 117983, NA, 154123, NA, NA, NA, 133275...  
## $ GNIS\_NAME <chr> NA, NA, "George Lake", NA, NA, "Emmit Wood Lake", "Emmit Wood Lake", N...  
## $ HUC2 <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 6, 3, 3, 3, 3, 3, 3, 11, 11, 8, 8, 11...  
## $ HUC8 <int> 3160112, 3160112, 3160109, 3150106, 3160201, 3160203, 3160203, 3140202...  
## $ LAKE\_ORIGIN <chr> "MAN\_MADE", "MAN\_MADE", "MAN\_MADE", "MAN\_MADE", "MAN\_MADE", "MAN\_MADE"...  
## $ LAKE\_ORIGIN12 <chr> "RESERVOIR", "RESERVOIR", "RESERVOIR", "RESERVOIR", "RESERVOIR", "RESE...  
## $ LAT\_DD83 <dbl> 33.29204, 33.29204, 34.22782, 33.76248, 32.51587, 31.62155, 31.62155, ...  
## $ LON\_DD83 <dbl> -87.39138, -87.39138, -86.84345, -86.05316, -87.86108, -88.35374, -88....  
## $ MAJ\_BASIN <chr> "3", "3", "3", "3", "3", "3", "3", "3", "3", "3", "3", "05\_06", "3", "...  
## $ MDCATY <chr> "NonREF07", "NonREF07", "NonREF07", "NonREF07", "NonREF07", "(20,50]",...  
## $ MISS\_BASIN <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "Yes...  
## $ NA\_L1CODE <int> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8...  
## $ NA\_L2CODE <dbl> 8.4, 8.4, 8.4, 8.4, 8.3, 8.3, 8.3, 8.3, 8.3, 8.4, 8.3, 8.4, 8.3, 8.3, ...  
## $ NA\_L3CODE <chr> "8.4.9", "8.4.9", "8.4.9", "8.4.1", "8.3.5", "8.3.5", "8.3.5", "8.3.5"...  
## $ NARS\_NAME <chr> "Holt Lock and Dam", "Holt Lock and Dam", "George Lake", NA, NA, "Emmi...  
## $ NES\_LAKE <chr> "NESLake", "NESLake", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...  
## $ NESLAKE\_ID <chr> "105", "105", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...  
## $ NESSTORET <chr> "105", "105", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...  
## $ OWN\_TYPE <chr> "Non Federal", "Non Federal", "Non Federal", "Non Federal", "Non Feder...  
## $ OWNSHP <chr> "Non-Federal", "Non-Federal", "Non-Federal", "Non-Federal", "Non-Feder...  
## $ PANEL <chr> "NLA07RVT2", "NLA07RVT2", "NLA07RVT", "NLA07RVT", "NLA07RVT", "NLA12RV...  
## $ PERIM\_KM <dbl> 104.2086662, 104.2086662, 11.8723053, 688.4591342, 698.1088203, 2.9989...  
## $ RCHCODE <chr> "3.16011E+12", "3.16011E+12", "3.16011E+12", "3.15011E+12", "3.16011E+...  
## $ RT\_NLA12 <chr> "T", "T", "S", "S", "S", "S", "S", NA, NA, NA, NA, NA, NA, "R", "S", "...  
## $ SITEID\_07 <chr> "NLA06608-2187", "NLA06608-2187", "NLA06608-0459", "NLA06608-2955", "N...  
## $ SITESAMP <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "N", "N", "N", "N", "N", "N", "Y", ...  
## $ SITETYPE <chr> "PROB", "PROB", "PROB", "PROB", "PROB", "PROB", "PROB", "PROB", "PROB"...  
## $ SIZE\_CLASS <chr> ">50", ">50", ">50", ">50", ">50", "(20,50]", "(20,50]", "(1,4]", "(10...  
## $ ST\_DSGN <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ ST\_MDCATY <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ ST\_PANEL <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ ST\_SITEID <chr> "NLA12\_AL-101", "NLA12\_AL-101", "NLA12\_AL-102", "NLA12\_AL-103", "NLA12...  
## $ ST\_STRATUM <chr> "AL\_NLA07RVT", "AL\_NLA07RVT", "AL\_NLA07RVT", "AL\_NLA07RVT", "AL\_NLA07R...  
## $ ST\_WGT\_ALL <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ ST\_WGT\_CAT <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ ST\_WGT\_DSGN <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ STATE <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL"...  
## $ STATE\_NLA <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL"...  
## $ STATE\_PCNT <int> 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, ...  
## $ STATECTY <int> 1073, 1073, 1043, 1055, 1063, 1129, 1129, 1031, 1085, 1009, 1053, 1071...  
## $ STATUS <chr> "Target\_Sampled", "Target\_Sampled", "Target\_Sampled", "Target\_Sampled"...  
## $ STATUS\_ST <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ STRATUM <chr> "AL\_NLA07RVT", "AL\_NLA07RVT", "AL\_NLA07RVT", "AL\_NLA07RVT", "AL\_NLA07R...  
## $ TNT <chr> "Target", "Target", "Target", "Target", "Target", "Target", "Target", ...  
## $ TNT\_ST <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ URBAN <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "No", "No", "No", "No", "No", "No",...  
## $ US\_L3CODE <int> 68, 68, 68, 67, 65, 65, 65, 65, 65, 68, 65, 68, 65, 65, 65, 67, 45, 65...  
## $ US\_L3CODE\_2015 <int> 68, 68, 68, 67, 65, 65, 65, 65, 65, 68, 65, 68, 65, 65, 65, 67, 45, 65...  
## $ US\_L4CODE <chr> "68f", "68f", "68d", "67g", "65b", "65f", "65f", "65d", "65a", "68d", ...  
## $ US\_L4CODE\_2015 <chr> "68f", "68f", "68d", "67g", "65p", "65f", "65f", "65d", "65a", "68d", ...  
## $ WGT\_ALL <dbl> 20.288885, 20.288885, 46.617902, 51.331899, 101.491046, 83.376963, 83....  
## $ WGT\_CAT <chr> "AL\_>4", "AL\_>4", "AL\_>4", "AL\_>4", "AL\_>4", "AL\_>4", "AL\_>4", "AL\_(1,...  
## $ WGT\_DSGN <dbl> 3.666670, 3.666670, 3.666667, 3.666667, 3.666667, 21.875000, 21.875000...  
## $ XCOORD <dbl> 795432.2, 795432.2, 835851.3, 913172.8, 759549.6, 721685.4, 721685.4, ...  
## $ YCOORD <dbl> -434840.8, -434840.8, -326047.1, -370517.5, -524957.0, -628108.3, -628...  
## $ ECO\_BIO <chr> "EHIGH", "EHIGH", "EHIGH", "EHIGH", "CPL", "CPL", "CPL", "CPL", "CPL",...  
## $ IM\_COMMENT <chr> "No phab done on this visit; visit 2 assigned as index visit", "Assign...  
## $ INDEX\_NLA <chr> "N", "Y", "Y", "Y", "Y", "N", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", ...  
## $ INDXSAMP\_BENT <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ INDXSAMP\_CHEM <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ INDXSAMP\_CHLA <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ INDXSAMP\_CORE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ INDXSAMP\_ENTE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ INDXSAMP\_MICR <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ INDXSAMP\_PHAB <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ INDXSAMP\_SDHG <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ INDXSAMP\_ZOOP <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ REF\_NLA12\_NUTR <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "R", NA, NA, "R", ...  
## $ RT\_NLA12\_BENT <chr> "T", "T", "S", "S", "S", "S", "S", NA, NA, NA, NA, NA, NA, "R", "S", "...  
## $ RT\_NLA12\_ZOOP <chr> "T", "T", "S", "S", "S", "S", "S", NA, NA, NA, NA, NA, NA, "R", "S", "...  
## $ SAMPLED\_BENT <chr> "Y", "Y", "Y", "N", "Y", "Y", "Y", NA, NA, NA, NA, NA, NA, "Y", "Y", "...  
## $ SAMPLED\_CORE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SAMPLED\_ENTE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SAMPLED\_MICR <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SAMPLED\_PHAB <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SAMPLED\_PHYT <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SAMPLED\_PROFILE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SAMPLED\_SDHG <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SAMPLED\_ZOCN <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", NA, NA, NA, NA, NA, NA, "Y", "Y", "...  
## $ SAMPLED\_ZOCR <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SAMPLED\_ZOFN <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", NA, NA, NA, NA, NA, NA, "Y", "Y", "...  
## $ SAMPLED\_ZOFR <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ SAMPLED\_ZOOP <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ PUBLICATION\_DATE <chr> "8/23/2016", "8/23/2016", "8/23/2016", "8/23/2016", "8/23/2016", "8/23...

Unfortunately, the RCHCODE values in the 2012 data that can be downloaded online were exported improperly, using scientific notation that cuts off the very important final digits. For example:

## # A tibble: 10 x 1  
## RCHCODE   
## <chr>   
## 1 8.01021E+12  
## 2 1.01401E+13  
## 3 1.01303E+13  
## 4 1.70401E+13  
## 5 1.025E+13   
## 6 1.60201E+13  
## 7 5.1402E+12   
## 8 3.0802E+12   
## 9 1.30202E+13  
## 10 6.04E+12

With luck, we can work around this issue.

## Filter

No need.

## Convert

No need.

## Select

nla2007 <- nla2007\_orig %>%  
 distinct(SITE\_ID, REACHCODE, COM\_ID)  
  
nla2012 <- nla2012\_orig %>%  
 distinct(SITE\_ID, SITEID\_07, COMID2012, COMID2007, COMIDS2007, GNIS\_ID, GNIS\_NAME, NESSTORET)

## Join to LAGOS\_US via NHDPlusV2 via the V1-to-V2 crosswalk.

We will use the nla2012 table as the left table in the join and also join through the original nhd\_plus dataset so that we can produce a list of NLA lakes that could not be linked to a lagoslakeid. The NLA lakes that don’t link will be assigned a lagoslakeid through our manual linking process (that starts with a spatial link).

Check on how many lakes didn’t link through each step.

**37** lakes out of **2664** NLA lakes cannot be linked to the NHDPlusV2. We did try the V1-to-V2 crosswalk and were not able to make any additional links. Far more lakes are unable to be linked to a lagoslakeid: **711**. This is due to the poor match between the NHDPlusV2 and the LAGOS Lake population. If we improve that link, these numbers will increase. This work is proposed.

## Select again

Select 2 columns from the NLA 2012 and lagoslakeid.

nla2012\_lagos <- nla2012\_lagos\_join %>%  
 select(  
 NLA2012\_SITE\_ID = SITE\_ID,  
 NLA2007\_SITE\_ID = SITEID\_07,  
 lagoslakeid  
 )

# Final, single crosswalk table

For the final table, we’ll join all the previous results together with the entire LAGOS-US lake population on the left side of the join. Then, in order to facilitate a user understanding of 1-to-many relationships in the table, we will generate some count columns that indicate when multiple identifiers in one dataset relate to a single identifier in another. Finally, we will do the final selection of columns for the output and rename a few more columns to match the naming template.

Glimpse at the final table.

## Observations: 569,329  
## Variables: 26  
## $ lagoslakeid <int> 31198, 31198, 31198, 31198, 31198, 31198, 122, 122, 12...  
## $ NHDHR\_PermanentIdentifier <chr> "120020498", "120020498", "120020498", "120020498", "1...  
## $ NHDHR\_ReachCode <chr> "04010201003330", "04010201003330", "04010201003330", ...  
## $ NHDHR\_AreaSqKm <dbl> 2.745000, 2.745000, 2.745000, 2.745000, 2.745000, 2.74...  
## $ GNIS\_ID <int> 649289, 649289, 649289, 649289, 649289, 649289, 653169...  
## $ GNIS\_Name <chr> "Perch Lake", "Perch Lake", "Perch Lake", "Perch Lake"...  
## $ LAGOS\_LakeName <chr> "Perch Lake", "Perch Lake", "Perch Lake", "Perch Lake"...  
## $ LAGOS\_CountyName <chr> "Carlton County", "Carlton County", "Carlton County", ...  
## $ LAGOS\_CountyFIPS <chr> "27017", "27017", "27017", "27017", "27017", "27017", ...  
## $ LAGOS\_LatitudeNAD83 <dbl> 46.68943, 46.68943, 46.68943, 46.68943, 46.68943, 46.6...  
## $ LAGOS\_LongitudeNAD83 <dbl> -92.67066, -92.67066, -92.67066, -92.67066, -92.67066,...  
## $ State <chr> "MN", "MN", "MN", "MN", "MN", "MN", "MN", "MN", "MN", ...  
## $ WQP\_MonitoringLocationIdentifier <chr> "FONDULAC-114A", "FONDULAC-114B", "FONDULAC\_WQX-114A",...  
## $ WQP\_MonitoringLocationName <chr> "Perch Lake (North Basin)", "Perch Lake (South Basin)"...  
## $ WQP\_ProviderName <chr> "STORET", "STORET", "STORET", "STORET", "STORET", "STO...  
## $ NHDPlusv2\_COMID <int> 1776072, 1776072, 1776072, 1776072, 1776072, 1776072, ...  
## $ NHDPlusv2\_ReachCode <chr> "04010201003330", "04010201003330", "04010201003330", ...  
## $ NHDPlusv2\_AreaSqKm <dbl> 2.720, 2.720, 2.720, 2.720, 2.720, 2.720, 1.348, 1.348...  
## $ LAGOSNE\_lagoslakeid <int> 31198, 31198, 31198, 31198, 31198, 31198, 122, 122, 12...  
## $ NLA2007\_SITE\_ID <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ NLA2012\_SITE\_ID <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ count\_wqp\_per\_lagos\_id <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 0, 0, 0, 0, 4, 4, ...  
## $ count\_nhdplusv2\_per\_lagos\_id <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 0, 1, 2, 2, ...  
## $ count\_lagos\_per\_nhdplusv2\_id <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, NA, 1, 1, 1,...  
## $ count\_lagosNE\_per\_lagos\_id <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ count\_lagos\_per\_lagosNE\_id <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...

An example of a group of lakes in the final table.

## # A tibble: 8 x 26  
## lagoslakeid NHDHR\_PermanentIdent~ NHDHR\_ReachCode NHDHR\_AreaSqKm GNIS\_ID  
## <int> <chr> <chr> <dbl> <int>  
## 1 10 132761914 01080206000814 1.13 608451  
## 2 10 132761914 01080206000814 1.13 608451  
## 3 10 132761914 01080206000814 1.13 608451  
## 4 9 132761873 01080206000810 0.186 604670  
## 5 9 132761873 01080206000810 0.186 604670  
## 6 9 132761873 01080206000810 0.186 604670  
## 7 33099 132761930 01080206000820 0.583 604600  
## 8 33099 132761930 01080206000820 0.583 604600  
## # ... with 21 more variables: GNIS\_Name <chr>, LAGOS\_LakeName <chr>,  
## # LAGOS\_CountyName <chr>, LAGOS\_CountyFIPS <chr>,  
## # LAGOS\_LatitudeNAD83 <dbl>, LAGOS\_LongitudeNAD83 <dbl>, State <chr>,  
## # WQP\_MonitoringLocationIdentifier <chr>,  
## # WQP\_MonitoringLocationName <chr>, WQP\_ProviderName <chr>,  
## # NHDPlusv2\_COMID <int>, NHDPlusv2\_ReachCode <chr>,  
## # NHDPlusv2\_AreaSqKm <dbl>, LAGOSNE\_lagoslakeid <int>,  
## # NLA2007\_SITE\_ID <chr>, NLA2012\_SITE\_ID <chr>,  
## # count\_wqp\_per\_lagos\_id <int>, count\_nhdplusv2\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_nhdplusv2\_id <int>, count\_lagosNE\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_lagosNE\_id <int>

Write out the final product.

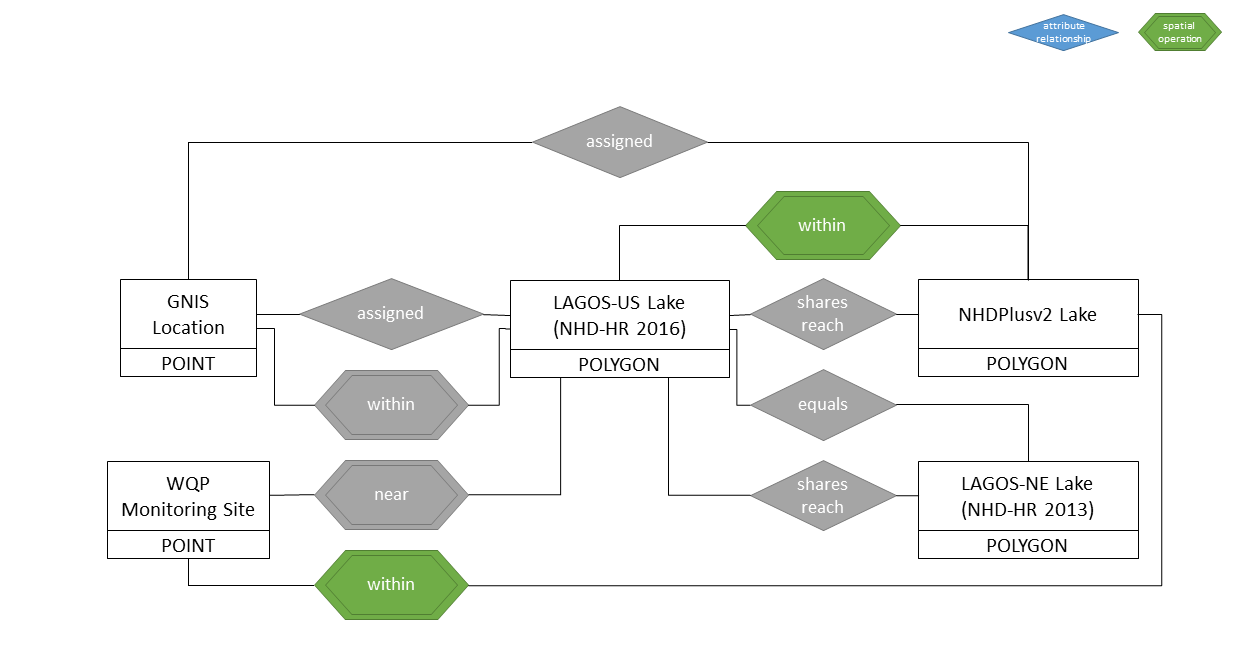
# List of to-dos and ideas to make more or better connections

1. PROBLEM: Almost 25% of lakes cannot be linked to their NHDPlusv2 identifier (and therefore their StreamCat/LakeCat identifier). SOLUTION: Link differing NHD version lakes through spatial analysis. This is the highest impact suggestion.
2. PROBLEM: There are still many unlinked WQP sites and some of them do link easily if adequate rules are proposed, but introduce a higher possibility of creating incorrect links for other similar sites. SOLUTION: We could make more links to WQP sites if we added a way to indicate “low confidence” links. For instance: we could add more sites that are within 100 m but have no name match.
3. PROBLEM: Some applications ask users to specify the lake they’re at when the user does not have GPS (Ewing and others harmful algal bloom app). SOLUTION: Add “nearest municipality” using TIGER data and a “closest” spatial join. (Discussed previously with Holly Ewing and was identified as desirable at this time, but cut from this draft for time.)
4. Manual links for LAGOS-NE lakes that have limno data (26 lakes) but that can’t be found through the Permanent\_Identifier or reach code in LAGOS-US. (In progress)
5. Manual links for any NLA sites that still don’t connect if we add the spatial analysis in TO-DO item (B).
6. Merge NHDPlusV2 names into the semi-colon list (cut for time–needs another join/distinct/verbose name calculation chunk)

# Things that can’t be done with the crosswalk

These items can be described as designed limitations because this lake identifier crosswalk is designed primarily to work with LAGOS. Or–if there is enough interest from the preview group–we can consider adding these capabilities. G) A user cannot find identifiers pertaining to a lake that isn’t in LAGOS-US. SOLUTION: Use outer joins and allow the crosswalk to grow a fair amount in size to allow NULLs in any of the identifiers. Revise the documentation to reflect the expansion. H) A user cannot connect WQP sites to the NHDPlusV2 representation of lakes. Example: 2 NHDPlusV2 lakes to 1 LAGOS-US lake. SOLUTION: Refine the entity-relationship model to join between the WQP and NHDPlusV2 directly. Rework subsequent joins in the script to accomodate the change. This change probably won’t affect the overall size or usability of the crosswalk too much.

## Proposed ER Diagram demonstrating change “A” and “H” above



*Proposed entity-relationship diagram showing two additions, in color*