# Assignment 1

Banner ID – B00924759

Email – viraj.joshi@dal.ca

## Part 2

Performed the following clean-up activity on the data files–

- Removed columns with all or more than half of the column values ‘blank’ or ‘NAN’. Preferred removing column over rows with those values to preserve data records.
- Removed rows in cases where certain columns had invalid values like ‘blanks’, ’Nan’ or ‘bad/inconsistent’ data.
- Removed first row from all datasets as it has meta data and not the data/column_names.

1. File - otnunit_aat_animals_8dc3_4d15_c278

i. Removed column ‘age’ as it had all NAN values.

ii. Removed column ‘lifestage’ with 3298 values blank.

iii. Cleaned up column ‘stock’ by introducing uniformity by ‘UNK’ to ‘UNKNOWN’

iv. Removed rows ‘length’ and ‘weight’ with NaN values and column ‘sex’ with all blank values.

2. File - otnunit_aat_datacenter_attributes_8a94_cefd_f8a3

i. Removed column ‘time_coverage_end’ and ‘time_coverage_start’ as it had all Null values.

ii. Removed record for datacenter ‘OTN-NEP’ as it is not referenced by other tables and has NaN values for datacenter_geospatial_lon_min, datacenter_geospatial_lon_max, datacenter_geospatial_lat_min, datacenter_geospatial_lat_max columns.

3. File - otnunit_aat_detections_9062_592

i. Removed columns ‘sensor_data’, ‘sensor_data_units’ and ‘detection_quality’ as they had more than half the values ‘blank’

ii. Removed columns ‘receiver_log_id’ , ‘depth’ , ‘uncertainty_in_latitude’, ‘depth_data_source’ and ‘uncertainty_in_longitude’, ‘uncertainty_in_depth’, ‘other_position_data’, ‘dataset_quality’ as they had all values ‘NaN’

iii. Cleaned records with inconsistent special character ‘?’ in key column.

4. File - otnunit_aat_manmade_platform_0735_7c9f_329c

i. Removed records with ‘platform_depth’ unknown.

ii. Certain key values were repeating. The difference was either ‘case-sensitivity’ or duplicity. Cleaned all such records.

5. File - otnunit_aat_project_attributes_f29c_fb21_23a3

i. Added ‘UNKNOWN' value in column ‘project_pi’ and ‘project_pi_contact’ for projects with blank value for this field. Could not remove the project records, as the project name was being referenced in other tables.

ii. Set ‘project_infourl’ for PRT, HFX, OBAS, V2LEOR2 to 'UNKNOWN'.

iii. Removed ‘project_doi’, ‘project_distribution_statement’ ,’ project_date_modified’ columns as it has NaN or Blank values.

iv. Performed clean-up for incorrect data in ‘geospatial_vertical_min’, ‘geospatial_vertical_max’,’ geospatial_vertical_positive’,’ time_coverage_start’,’ time_coverage_end’.

6. File - otnunit_aat_receivers_c595_05f4_68b2

i. Removed ‘frequencies_monitored’, ‘receiver_coding_scheme’,

ii. Removed rows for receivers where ‘receiver_reference_id’ was blank.

iii. Removed columns ‘bottom_depth’ with 4284 blank values, depth with 4956 blank values and ‘deployment_comments’ with 15522 blank values.

iv. Removed 6 rows where ‘receiver_serial_number’ was ‘unknown’ and 2 rows where it was ‘-‘. Set ‘receiver_mfg’ to "UNKNOWN" for 795 rows as deleting rows for blank data would mean losing receiver data. Removed two rows which had blank ‘serial numbers’.

v. Set datetime to ‘1900-01-01’ for blank values.

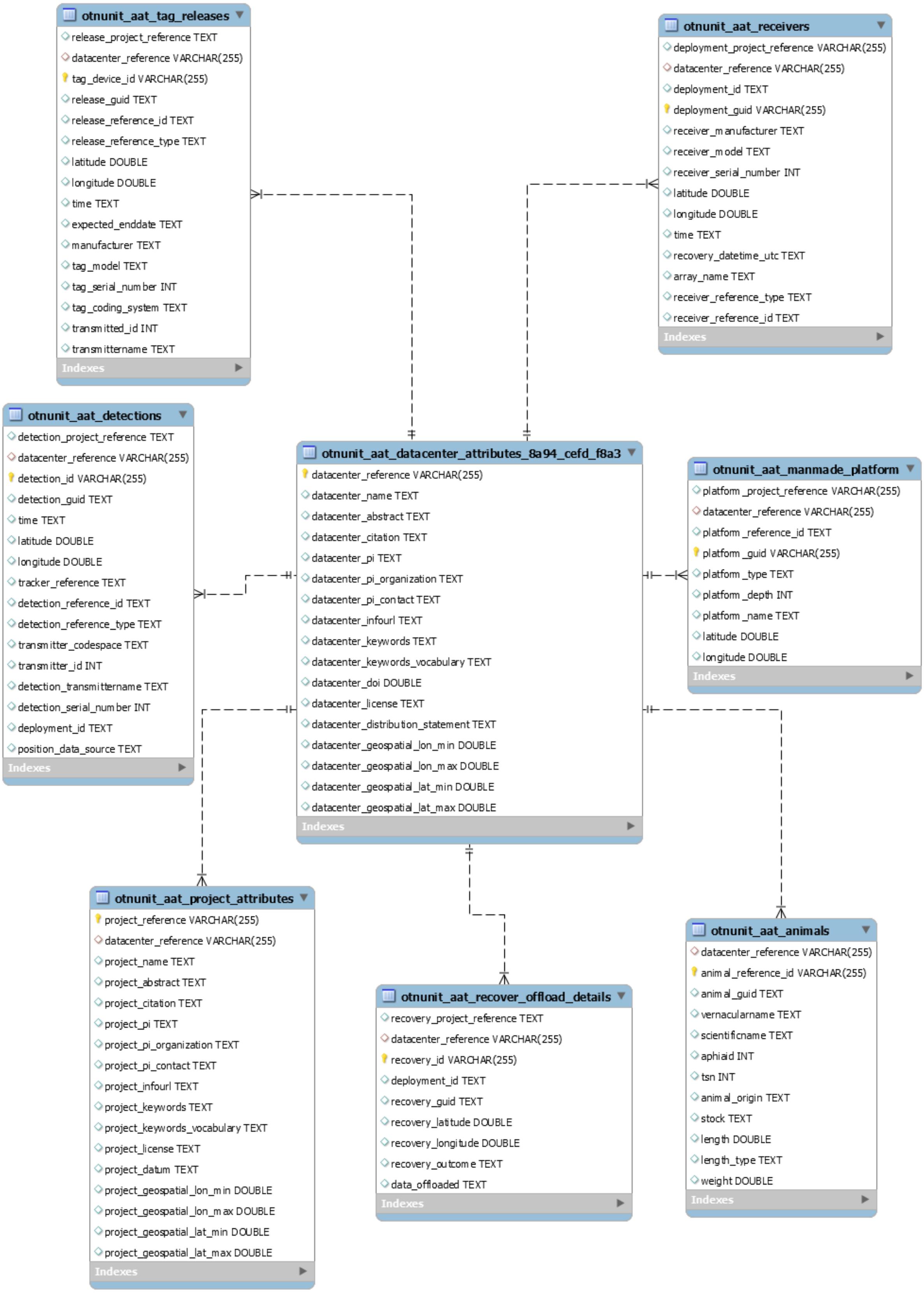
7. File - otnunit_aat_recover_offload_details

i. Removed column ‘recovery_datetime_utc’ , ‘offload_datetime_utc’, ‘log_filenames’ , ‘clock_synchronized’, ‘recovered_by’ which were all blank and ‘recovery_latitude’ with NaN values.

8. File – otnunit_aat_tag_releases_b793_03e7_a230

i. Removed column ‘tag frequency’, ‘transmitter_type’, ‘tag_programming_id’ with all blank values.

# Before Normalization



otnunit_aat_tag_releases
release_project_reference TEXT
datacenter_reference VARCHAR(255)
tag_device_id VARCHAR(255)
release_guid TEXT
release_reference_id TEXT
release_reference_type TEXT
latitude DOUBLE
longitude DOUBLE
time TEXT
expected_enddate TEXT
manufacturer TEXT
tag_model TEXT
tag_serial_number INT
tag_coding_system TEXT
transmitted_id INT
transmittername TEXT
Indexes

otnunit_aat_receivers
deployment_project_reference VARCHAR(255)
datacenter_reference VARCHAR(255)
deployment_id TEXT
deployment_guid VARCHAR(255)
receiver_manufacturer TEXT
receiver_model TEXT
receiver_serial_number INT
latitude DOUBLE
longitude DOUBLE
time TEXT
recovery_datetime_utc TEXT
array_name TEXT
receiver_reference_type TEXT
receiver_reference_id TEXT
Indexes

otnunit_aat_detections
detection_project_reference TEXT
datacenter_reference VARCHAR(255)
detection_id VARCHAR(255)
detection_guid TEXT
time TEXT
latitude DOUBLE
longitude DOUBLE
tracker_reference TEXT
detection_reference_id TEXT
detection_reference_type TEXT
transmitter_codespace TEXT
transmitter_id INT
detection_transmittername TEXT
detection_serial_number INT
deployment_id TEXT
position_data_source TEXT
Indexes

otnunit_aat_datacenter_attributes_8a94_cefd_f8a3
datacenter_reference VARCHAR(255)
datacenter_name TEXT
datacenter_abstract TEXT
datacenter_citation TEXT
datacenter_pi TEXT
datacenter_pi_organization TEXT
datacenter_pi_contact TEXT
datacenter_infourl TEXT
datacenter_keywords TEXT
datacenter_keywords_vocabulary TEXT
datacenter_doi DOUBLE
datacenter_license TEXT
datacenter_distribution_statement TEXT
datacenter_geospatial_lon_min DOUBLE
datacenter_geospatial_lon_max DOUBLE
datacenter_geospatial_lat_min DOUBLE
datacenter_geospatial_lat_max DOUBLE
Indexes

otnunit_aat_manmade_platform
platform_project_reference VARCHAR(255)
datacenter_reference VARCHAR(255)
platform_reference_id TEXT
platform_guid VARCHAR(255)
platform_type TEXT
platform_depth INT
platform_name TEXT
latitude DOUBLE
longitude DOUBLE
Indexes

otnunit_aat_project_attributes
project_reference VARCHAR(255)
datacenter_reference VARCHAR(255)
project_name TEXT
project_abstract TEXT
project_citation TEXT
project_pi TEXT
project_pi_organization TEXT
project_pi_contact TEXT
project_infourl TEXT
project_keywords TEXT
project_keywords_vocabulary TEXT
project_license TEXT
project_datum TEXT
project_geospatial_lon_min DOUBLE
project_geospatial_lon_max DOUBLE
project_geospatial_lat_min DOUBLE
project_geospatial_lat_max DOUBLE
Indexes

otnunit_aat_recover_offload_details
recovery_project_reference TEXT
datacenter_reference VARCHAR(255)
recovery_id VARCHAR(255)
deployment_id TEXT
recovery_guid TEXT
recovery_latitude DOUBLE
recovery_longitude DOUBLE
recovery_outcome TEXT
data_offloaded TEXT
Indexes

otnunit_aat_animals
datacenter_reference VARCHAR(255)
animal_reference_id VARCHAR(255)
animal_guid TEXT
vernacularname TEXT
scientificname TEXT
aphiaid INT
tsn INT
animal_origin TEXT
stock TEXT
length DOUBLE
length_type TEXT
weight DOUBLE
Indexes

# Normalization

## 1. Table - otnunit_aat_animals

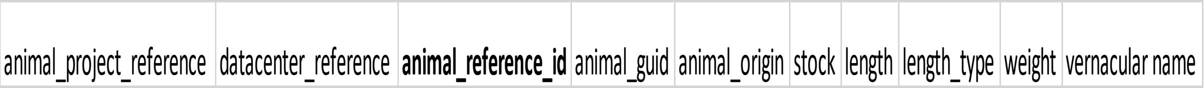
i. We observe that the data values in this table are atomic. Hence, the table is in 1NF.

But, if we notice, non-key columns columns ‘scientificname’,’aphiaid’,’tsn’ and ‘animal_origin’ can be determined by the non-key column ‘vernacular_name’.

We normalize this table by creating the following tables–

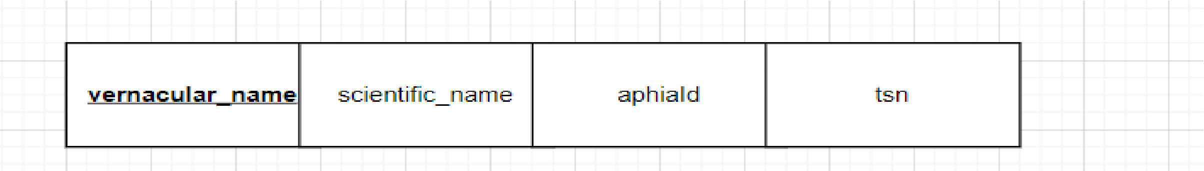
**otnunit_aat_animals –**

‘vernacular_name’ of otnunit_aat_animals table refers primary key – ‘vernacular_name’ of animal_type_data table.



| animal_project_reference | datacenter_reference | **animal_reference_id** | animal_guid | animal_origin | stock | length | length_type | weight | vernacular name |
|---|---|---|---|---|---|---|---|---|---|

**animal_type_data -**



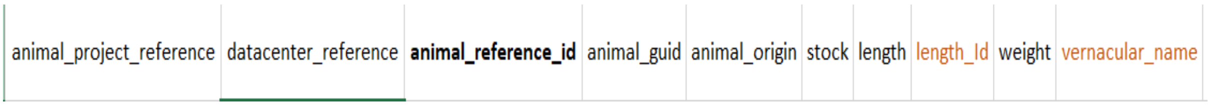
| vernacular_name | scientific_name | aphiaId | tsn |
|---|---|---|---|

ii. Above table shows that ‘length_type’ is functionally dependent on ‘length’. If we remove records for a certain length, we may lose critical information about the length type as well. So, we create a new table ‘**length_type_data**’ with ‘length_Id’ as the primary key and ‘length_type’ column. This way we can preserve length type values, even if a length record is removed from **otnunit_aat_animals.**

We achieve 3NF by creating the following tables –

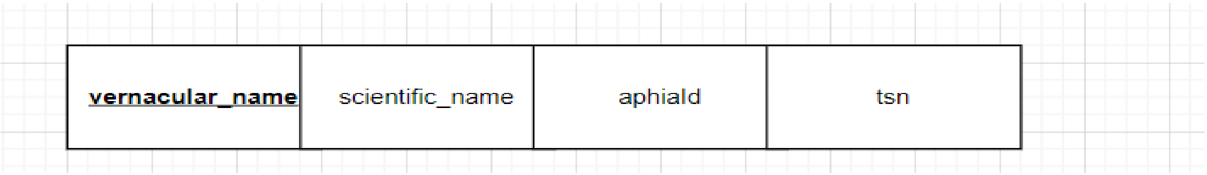
**otnunit_aat_animals -**

Foreign key ‘length_Id’ references ‘length_Id’ column of ‘**length_type_data’** table



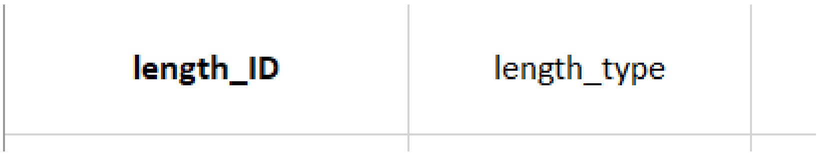
| animal_project_reference | datacenter_reference | **animal_reference_id** | animal_guid | animal_origin | stock | length | length_Id | weight | vernacular_name |
|---|---|---|---|---|---|---|---|---|---|

**animal_type_data**



| **vernacular_name** | scientific_name | aphiaId | tsn |
|---|---|---|---|

**length_type_data**



| **length_ID** | length_type |
|---|---|

## 2. Table - otnunit_aat_detections_9062

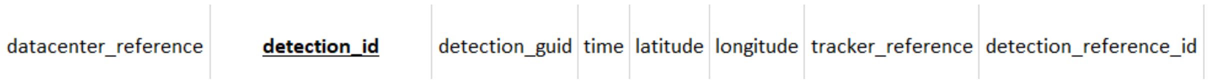
i. This table is already under 1NF.

But we observe that non-key columns ‘transmitter_codespace’ and ’detection_transmittername’ can be determined by non-key column ‘transmitter_ID’. Also, ‘detection_transmittername’ is a combination of transmitter_codespace + transmitter_ID’.

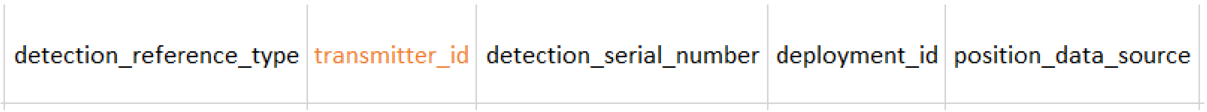
So, we normalize the table by creating the following –

**otnunit_aat_detections_9062 –**

Foreign key ‘transmitter_id’ refers ‘transamitter_id’ of ‘**transmitter_data**’ table

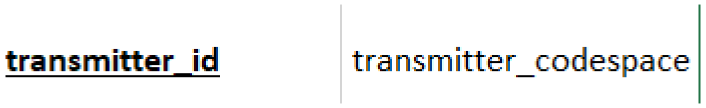


| datacenter_reference | **detection_id** | detection_guid | time | latitude | longitude | tracker_reference | detection_reference_id |
|---|---|---|---|---|---|---|---|



| detection_reference_type | transmitter_id | detection_serial_number | deployment_id | position_data_source |
|---|---|---|---|---|

**transmitter_data –**



| **transmitter_id** | transmitter_codespace |
|---|---|

## 3. Table – otnunit_aat_tag_releases

i. The table is in 1NF as all values at atomic.

ii. a. We observe that, non-key attribute ‘tag_coding_system’, ‘transmitter_name’ is dependent on non-key attribute ‘transmitter_id’. Also, ‘transmitter_name’ is a combination of ‘tag_coding_system’ + ‘transmitter_name’.

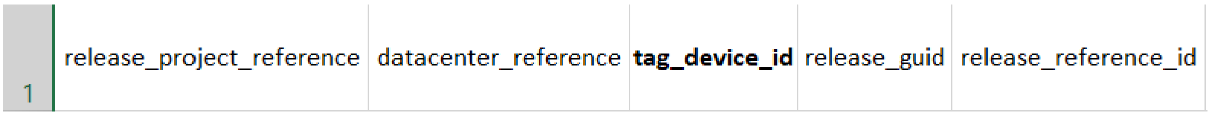
We have **already** created transmitter_data table for the previous normalization which had columns ‘transmitter_id’ and ‘transmitter_codespace’ where ‘transmitter_codespace’ has similar values as ‘tag-coding-system’.

ii. b. We observe that, non-key attribute ‘release_reference_type’ depends upon non-key attribute ‘release_reference_id’.

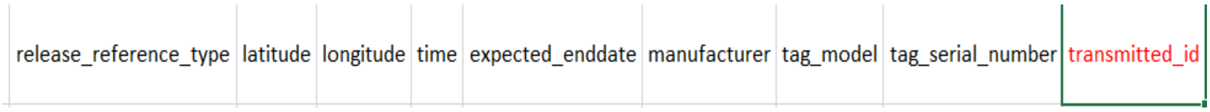
We normalize to 3$^{rd}$ NF by having the following tables –

**otnunit_aat_tag_releases –**

Foreign-key ‘transmitter_id’ refers ‘transmitter_id’ of **transmitter_data** table.

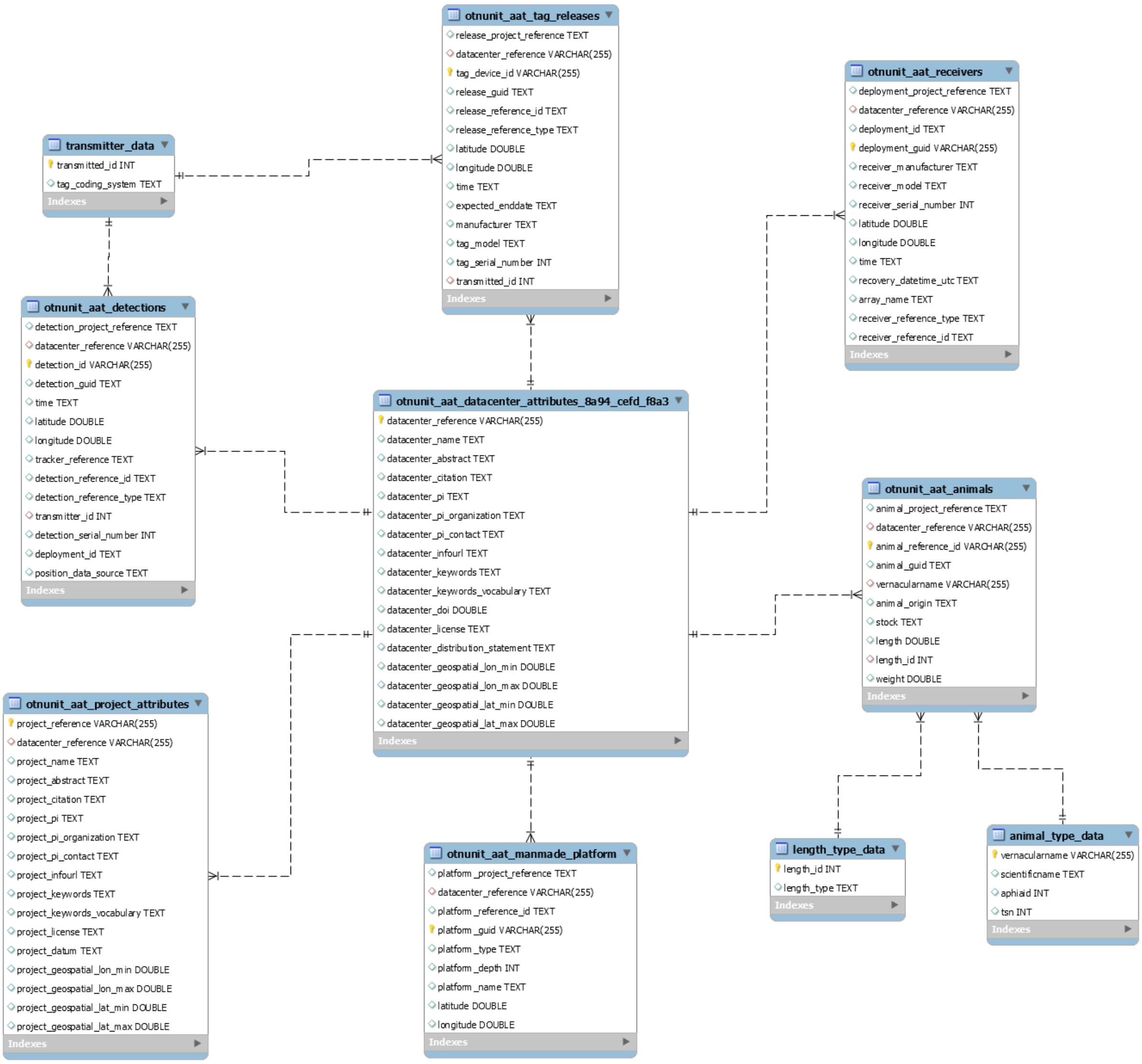


| 1 | release_project_reference | datacenter_reference | **tag_device_id** | release_guid | release_reference_id |
|---|---|---|---|---|---|



| release_reference_type | latitude | longitude | time | expected_enddate | manufacturer | tag_model | tag_serial_number | transmitted_id |
|---|---|---|---|---|---|---|---|---|

# After Normalization


otnunit_aat_tag_releases
release_project_reference TEXT
datacenter_reference VARCHAR(255)
tag_device_id VARCHAR(255)
release_guid TEXT
release_reference_id TEXT
release_reference_type TEXT
latitude DOUBLE
longitude DOUBLE
time TEXT
expected_enddate TEXT
manufacturer TEXT
tag_model TEXT
tag_serial_number INT
transmitted_id INT
Indexes
otnunit_aat_receivers
deployment_project_reference TEXT
datacenter_reference VARCHAR(255)
deployment_id TEXT
deployment_guid VARCHAR(255)
receiver_manufacturer TEXT
receiver_model TEXT
receiver_serial_number INT
latitude DOUBLE
longitude DOUBLE
time TEXT
recovery_datetime_utc TEXT
array_name TEXT
receiver_reference_type TEXT
receiver_reference_id TEXT
Indexes
transmitter_data
transmitted_id INT
tag_coding_system TEXT
Indexes
otnunit_aat_detections
detection_project_reference TEXT
datacenter_reference VARCHAR(255)
detection_id VARCHAR(255)
detection_guid TEXT
time TEXT
latitude DOUBLE
longitude DOUBLE
tracker_reference TEXT
detection_reference_id TEXT
detection_reference_type TEXT
transmitted_id INT
detection_serial_number INT
deployment_id TEXT
position_data_source TEXT
Indexes
otnunit_aat_datacenter_attributes_8a94_cefd_f8a3
datacenter_reference VARCHAR(255)
datacenter_name TEXT
datacenter_abstract TEXT
datacenter_citation TEXT
datacenter_pi TEXT
datacenter_pi_organization TEXT
datacenter_pi_contact TEXT
datacenter_infourl TEXT
datacenter_keywords TEXT
datacenter_keywords_vocabulary TEXT
datacenter_doi DOUBLE
datacenter_license TEXT
datacenter_distribution_statement TEXT
datacenter_geospatial_lon_min DOUBLE
datacenter_geospatial_lon_max DOUBLE
datacenter_geospatial_lat_min DOUBLE
datacenter_geospatial_lat_max DOUBLE
Indexes
otnunit_aat_animals
animal_project_reference TEXT
datacenter_reference VARCHAR(255)
animal_reference_id VARCHAR(255)
animal_guid TEXT
vernacularname VARCHAR(255)
animal_origin TEXT
stock TEXT
length DOUBLE
length_id INT
weight DOUBLE
Indexes
otnunit_aat_project_attributes
project_reference VARCHAR(255)
datacenter_reference VARCHAR(255)
project_name TEXT
project_abstract TEXT
project_citation TEXT
project_pi TEXT
project_pi_organization TEXT
project_pi_contact TEXT
project_infourl TEXT
project_keywords TEXT
project_keywords_vocabulary TEXT
project_license TEXT
project_datum TEXT
project_geospatial_lon_min DOUBLE
project_geospatial_lon_max DOUBLE
project_geospatial_lat_min DOUBLE
project_geospatial_lat_max DOUBLE
Indexes
otnunit_aat_manmade_platform
platform_project_reference TEXT
datacenter_reference VARCHAR(255)
platform_reference_id TEXT
platform_guid VARCHAR(255)
platform_type TEXT
platform_depth INT
platform_name TEXT
latitude DOUBLE
longitude DOUBLE
Indexes
length_type_data
length_id INT
length_type TEXT
Indexes
animal_type_data
vernacularname VARCHAR(255)
scientificname TEXT
aphiaid INT
tsn INT
Indexes