CSCI5408
ASSIGNMENT 1 – PROBLEM 2

Format Ocean Tracking Data and Report

Dhrumil Amish Shah
B00857606

Problem #2: Format Ocean Tracking Data and Report

Datasets and attributes discovered

Different datasets and attributes related to Ocean Tracking Network(OTN) taken from the website https://oceantrackingnetwork.org/about/#oceanmonitoring are:

1. **Marine Departments** is a dataset with attributes like department id, name, description, manager id, manager joining date.
2. **Marine Scientists** is a dataset with attributes like unique id, name, birthdate, address, email, contact, joining date, leaving date, department id, supervisor id.
3. **OTN Council** is a dataset with attributes like council member id, member name, designation, organization, email, contact, department, is voting or non-voting council member and other necessary attributes, joining date.
   Similarly, we can have a dataset of OTN ISAC, OTN SAC, OTN IDMC. Also, we can group all the members in one dataset and add an attribute to identify which committee or council a member belongs to.
4. **Funding** is a dataset with attributes like provider name, funding amount, approve date, given date, funding period, funding description, provider contact, provider email address, provider address.
5. **Aquatic Species Classes** is a dataset with attributes like species id, name, description, ecosystem, lifespan, type, habitat, size, gender.
6. **Acoustic Tags** is a dataset with attributes like unique id, unique code, attachment type, attached to species, date of use, manufacturing date and company, frequency, lifespan.
7. **Acoustic Receivers** is a dataset with attributes like receiver id, receiver location, receiver category, price, model, manufacturer and all the data received by different aquatic species.
8. **VEMCO Mobile Transceivers (VMTs)** is a dataset with attributes like unique id, location updates, attached to species, size, price, model. VMT works as both transmitter and receiver.
9. **Wave Gliders** is a dataset that collects oceanographic and weather data. Other attributes are unique id, location updates, model, manufacturer, data received by different aquatic species.
10. **Slocum Gliders** is a dataset that collects waves information. Other attributes are unique id, location updates, model, manufacturer, data received by different aquatic species.
11. **Data Center** is a dataset that collects information about the data centers in OTN. The associated attributes are data center number, data center name, data center location, data center description.

Cleaning/Transforming dataset

1. **otnunit_aat_animals_8dc3_4d15_c278.csv (Total records - 3809)**
   1. The first row is empty. Hence it is removed completely.
   2. The column **animal_guid** is a composite attribute that is a combination of columns **datacenter_reference**, **animal_project_reference** and **animal_reference_id**. Hence, this column is removed completely as it can be derived from the other three columns.
   3. The column **taxonrank** has not recorded data for any entry in the dataset. It doesn't derive any value in the dataset. Thus I removed the **taxonrank** column completely.
   4. The **animal_origin** column can be inferred from columns **animal_project_reference**, **datacenter_reference**, **vernacularname**, **scientificname**, **aphiaid**, and **tsn**. I added the value "W" in all the entries for the **animal_origin** column where the column **vernacularname** has the value sevengill shark (12 replacements).
   5. Out of 3809 records, the **stock** column has 323 unknowns. The **stock** column has two values for unknown (i.e., UNK and UNKNOWN). I kept UNK for consistency and, I replaced UNKNOWN with UNK. Values in total 99 cells are replaced from UNKNOWN to UNK.
   6. The blank entries in the **stock** column can be inferred from **animal_project_reference**, **datacenter_reference**, **vernacularname**, **scientificname**, **aphiaid**, **tsn**, **animal_origin** and

length_type. All the **stock** associated with the value blue shark has value NW Atlantic. Thus for blank entries, I added the value NW Atlantic. Values in total 22 cells are replaced from blank to NW Atlantic.

7. For the rest of the **vernicularname**, the value in the **stock** column is unknown. I added UNK where entries are blank. Values in total 141 cells are replaced from blank to UNK.
8. Not a Number(NaN) as a value in the **length** column is not allowed. In the **length** column, we can either have empty cells or -1 or NULL instead of NaN values. I replaced NaN values with NULL. Values in total 118 cells are replaced from NaN to NULL.
9. There is a total of 116 blank cells in column **length_type**. I filled them with UNK.
10. I followed the same process for columns **weight** and **age**. For the column **weight**, values in total 737 cells are replaced from NaN to NULL. Similarly, for column **age**, values in total 3659 cells are replaced from NaN to NULL.
11. The **age** column has 3659 unknowns out of 3809 records. (i.e., 96.06% of data is missing). It has no strong involvement in the analysis but I kept the column as it can be filled once the data is available.
12. Similarly, for columns **life_stage** and **sex**, 3298 and 3664 records are unknown out of 3809 records. (i.e. 86.58% and 96.19% of data are missing respectively). Data can be filled in once available. For now, I filled UNK in **life_stage** and replaced blank with U in column **sex**.

2. **otnunit_aat_datacenter_attributes_8a94_cefd_f8a3.csv(Total records - 4)**
    1. The first row specifies that **time_coverage_start** and **time_coverage_end** are in UTC. These values are not data but data types for the mentioned columns. Hence, this record is deleted.
    2. The columns, **datacenter_distribution_statement**, **datacenter_date_modified, time_coverage_start** and **time_coverage_end** are completely empty. Hence, I removed these columns from the dataset. These columns can be added back when ample information is available.
    3. The column **datacenter_license** and **datacenter_abstract** had 3 rows which were inconsistent as they contained special characters. The special characters were removed to make the data in those columns uniform.
    4. There were NaN values in 1 row in columns **datacenter_geospatial_lon_min**, **datacenter_geospatial_lon_max**, **datacenter_geospatial_lat_min** and **datacenter_geospatial_lat_max**. The NaN value was replaced with NULL for columns **datacenter_geospatial_lon_min**, **datacenter_geospatial_lon_max**, **datacenter_geospatial_lat_min** and **datacenter_geospatial_lat_max**
    as the valid values for these columns would always be number.
    5. The columns **datacenter_abstract**, **datacenter_pi**, **datacenter_pi_organization**, **datacenter_pi_contact**, **datacenter_infourl**, **datacenter_keywords datacenter_keywords_vocabulary**, **datacenter_doi** and **datacenter_license** have same value for all the records.

3. **otnunit_aat_detections_9062_5923_1394.csv(Total records - 218978)**
    1. The first row specifies that **latitude** and **longitude** are in degrees_north and degrees_east respectively. The values for column **depth** and **time** are in m and UTC. These values are not data but data types for the mentioned columns. Hence, this record is deleted.
    2. The column **detection_guid** is a composite attribute that is a combination of columns **detection_project_reference**, **datacenter_reference** and **detection_id**. Hence, this column is removed completely as it can be derived from the other three columns.

3. The column **detection_transmittername** is a composite attribute that is a combination of columns **transmitter_codespace** and **transmitter_id**. Hence, this column is removed completely as it can be derived from the other two columns.
4. The column **sensor_data** had 215274 out of 218978 cells (i.e., 98% cells) empty i.e. no values. In Ocean Tracking and monitoring, sensor data is one of the crucial factors, thus I retained the column but I changed the blank cells to NULL values.
5. The column **sensor_data_units** had 215110 out of 218978 rows empty which I updated with NA.
6. The column **receiver_log_id** was empty completely, thus it was removed.
7. Around 97.69% of the column data for **detection_quality** (about 213924 rows) were empty. Hence, it was removed completely.
8. There were only NaN values in columns **depth**, **uncertainty_in_latitude**, **uncertainty_in_longitude**. Hence, they were removed to use the database space efficiently.
9. There were no values in columns **depth_data_source**, **uncertainty_in_depth**, **other_position_data** and **dataset_quality**. Thus, they were removed completely.

4. otnunit_aat_manmade_platform_0735_7c9f_329c.csv (Total records – 8938)

1. The first row specifies that **latitude** and **longitude** are in degrees_north and degrees_east respectively. These values are not data but data types for the mentioned columns. Hence, this record is deleted.
2. The column **platform_guid** is a composite attribute that is a combination of columns **datacenter_reference**, **platform_project_reference** and **platform_reference_id**. Hence, this column is removed completely as it can be derived from the other three columns.
3. The column **platform_depth** had 2261 out of 8938 cells (i.e., 25.30% cells) filled with NaN values. Depth cannot have non-numeric values. Hence, these cells were replaced with NULL.
4. The columns **latitude** and **longitude** had 10 cells with the value NaN. These cells were replaced with NULL for consistency.
5. The **platform_reference_id** and **platform_name** columns are the same. There is no point in having the same data in two different columns. Hence, I removed the **platform_name** column from the dataset.

5. otnunit_aat_project_attributes_f29c_fb21_23a3 (Total records - 300)

1. The first row specifies that **project_geospatial_lon_min** and **project_geospatial_lon_max** are in degrees_east and **project_geospatial_lat_min** and **project_geospatial_lat_max** are in degrees_north. These values are not data but data types for the mentioned columns. Hence, this row was deleted.
2. The columns **project_references**, **project_doi**, **project_distribution_statement**, **project_date_modified**, **project_linestring**, **geospatial_vertical_positive**, **time_coverage_start** and **time_coverage_end** are completely empty. Hence, I removed these columns from the dataset. These columns can be added back when ample information is available.
3. The attribute **project_abstract** had 3 blank cells which I filled with NA.
4. The column **project_citation** had 12 blank cells which I filled with NA.
5. The column **project_poi** had 16 blank cells which I filled with NA.
6. The column **project_infourl** had 1 value <NULL>, 2 values blank and other NA values. They were all replaced with NA for consistency.
7. The columns **project_keywords_vocabulary**, **project_license** and **project_datum** have the same values for all the records.
8. The blanks were replaced with NULL as it is for columns **geospatial_vertical_min** and **geospatial_vertical_max**.

6. **otnunit_aat_receivers_c595_05f4_68b2.csv (Total records – 18786)**
   1. The first row specifies that **latitude** and **longitude** are in degrees_north and degrees_east respectively, **time** and **recovery_datetime_utc** are in UTC and **bottom_depth** and **depth** are in m(meters). These values are not data but data types for the mentioned columns. Hence, this row was deleted.
   2. The column **deployment_guid** is a composite attribute that is a combination of columns **datacenter_reference**, **deployment_project_reference**, **deployment_id**. Hence, this column is removed completely as it can be derived from the other three columns.
   3. The three columns, namely, **frequencies_monitored**, **receiver_coding_scheme** and **deployed_by** are completely empty. Hence, I removed these columns from the dataset. These columns can be added back when ample information is available.
   4. The column **expected_receiver_life** is populated with NaN. This column is deleted because it provides no information.
   5. There was a total of 800 records empty in the column **receiver_manufacturer** for which 635 records belonged to the VR2W model. For the VR2W model specifically, 10953 records had value VEMCO as **receiver_manufacturer** and VR2W as **receiver_model**. Thus, I filled VEMCO as the **receiver_manufacturer** for the 635 blank cells against VR2W. The remaining 165 records out of 800 records identified earlier were replaced with UNK as their **receiver_manufacturer** for the corresponding **receiver_model** was not available.
   6. For the column **deployment_comments** 15525 out of 18786 cells were empty(i.e., 82.64% empty values). Filled empty cells with the UNK flag.
   7. The column **receiver_serial_number** had 2 blank cells. Filled the empty cells in column **receiver_serial_number** with value UNK for consistency. Also, the column **recovery_datetime_utc** had 3409 blank cells. Filled the empty cells in column **recovery_datetime_utc** with the value 0000-00-00T00:00:00Z.
   8. The columns **bottom_depth** and depth consist of 4286 and 4956 NaN values which were replaced with NULL.

7. **otnunit_aat_recover_offload_details_4b23_f002_f89a.csv (Total records – 37203)**
   1. The first row is empty other than value UTC in columns **recovery_datetime_utc** and **offload_datetime_utc**. Hence, this row is deleted.
   2. The column **recovery_guide** is a composite attribute that is a combination of columns **datacenter_reference**, **recovery_id**, **deployment_id**. Hence, this column is removed completely as it can be derived from the other three columns.
   3. The columns **clock_synchronized** and **recovered_by** are completely empty. Hence, I removed both the columns from the dataset. These columns can be added back when ample information is available.
   4. The percentage of data unavailable for columns **recovery_datetime_utc**, **offload_datetime_utc**, **log_filenames** and **recovery_comments** are 58.67% (21,826 from 37203), 59.7% (22210 from 37203), 52.13% (19393 from 37203) and 82.31% (30624 from 37203). I kept these columns in the dataset and fill the empty cells as below:
      i. 0000-00-00T00:00:00Z in **recovery_datetime_utc** and **offload_datetime_utc**.
      ii. UNK in **log_filenames** and **recovery_comments**.

8. **otnunit_aat_tag_releases_b793_03e7_a230.csv (Total records – 3837)**
   1. The first row specifies that **latitude** and **longitude** are in degrees_north and degrees_east respectively and **time** and **expected_enddate** are in UTC. These values are not data but data types for the mentioned columns. Hence, this row is deleted.

2. The column **release_guid** is a composite attribute that is a combination of columns **datacenter_reference**, **release_project_reference**, **tag_device_id**. Hence, this column is removed completely as it can be derived from the other three columns.
3. The column **transmittername** is a composite attribute that is a combination of columns **tag_coding_system** and **transmitted_id.** Hence, this column is removed completely as it can be derived from the other two columns.
4. The columns, namely, **tag_frequency**, **transmitter_type** and **tag_programming_id** are completely empty. Hence, I removed these columns from the dataset. These columns can be added back when ample information is available.

Normalization Logic

1. **otnunit_aat_animals**
   - This dataset was already in 1NF.
   - The columns vernacularname, scientificname, aphiaid and tsn were added in a separate dataset called otnunit_aat_animals_species. The column vernacularname acts as a foreign key in otnunit_aat_animals.

2. **otnunit_aat_detections**
   - This dataset was already in 1NF.
   - The columns transmitter_codespan and transmitter_id were added in a separate dataset called otnunit_aat_transmitter_data. The column transmitter_id acts as a foreign key in otnunit_aat_detections.

3. **otnunit_aat_tag_releases**
   - This dataset was already in 1NF.
   - The columns release_reference_id, release_reference_type and release_project_reference were added in a separate dataset called otnunit_aat_tag_releases_details. The column release_reference_id acts as a foreign key in otn_aat_tag_releases.
   - The columns tag_device_id, tag_model, tag_serial_number and tag_coding_system were added in a separate dataset called otn_aat_tag_data. The column tag_device_id acts as a foreign key in otn_tag_releases.

4. **otnunit_aat_receivers**
   - This dataset was already in 1NF.
   - The columns deployment_id, receiver_serial_number, receiver_reference_type, receiver_reference_id, receiver_manufacturer_model, receiver_model were added in a separate dataset called otnunit_aat_receivers_details. It is in a one-to-one relationship with otnunit_aat_receivers.

⇨ **Figure 1** is the ERD generated using MySQL Workbench before any normalization. Since there was a total of eight datasets, there are a total of eight entities.

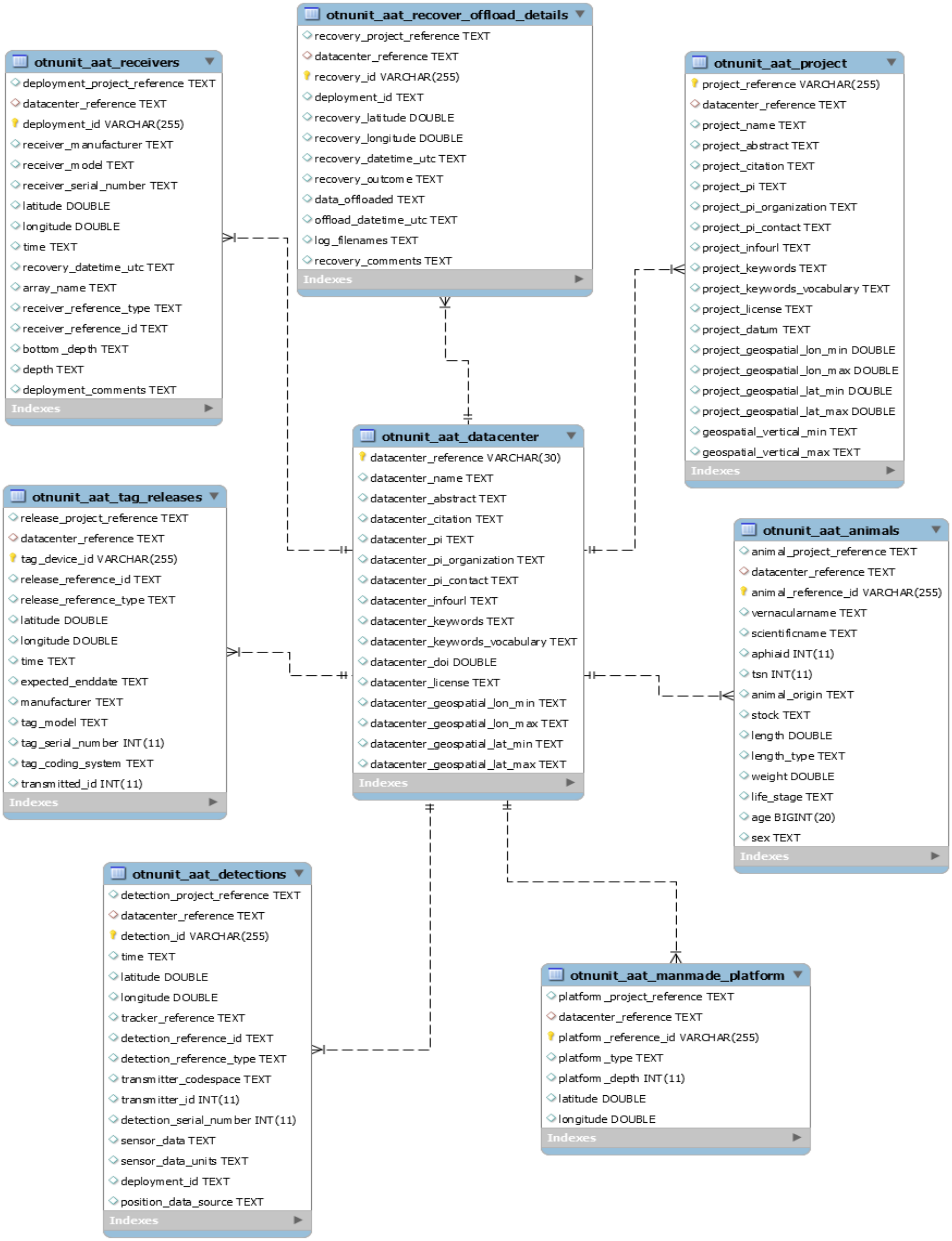
otnunit_aat_receivers
deployment_project_reference TEXT
datacenter_reference TEXT
deployment_id VARCHAR(255)
receiver_manufacturer TEXT
receiver_model TEXT
receiver_serial_number TEXT
latitude DOUBLE
longitude DOUBLE
time TEXT
recovery_datetime_utc TEXT
array_name TEXT
receiver_reference_type TEXT
receiver_reference_id TEXT
bottom_depth TEXT
depth TEXT
deployment_comments TEXT
Indexes
otnunit_aat_recover_offload_details
recovery_project_reference TEXT
datacenter_reference TEXT
recovery_id VARCHAR(255)
deployment_id TEXT
recovery_latitude DOUBLE
recovery_longitude DOUBLE
recovery_datetime_utc TEXT
recovery_outcome TEXT
data_offloaded TEXT
offload_datetime_utc TEXT
log_filenames TEXT
recovery_comments TEXT
Indexes
otnunit_aat_project
project_reference VARCHAR(255)
datacenter_reference TEXT
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
geospatial_vertical_min TEXT
geospatial_vertical_max TEXT
Indexes
otnunit_aat_datacenter
datacenter_reference VARCHAR(30)
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_geospatial_lon_min TEXT
datacenter_geospatial_lon_max TEXT
datacenter_geospatial_lat_min TEXT
datacenter_geospatial_lat_max TEXT
Indexes
otnunit_aat_tag_releases
release_project_reference TEXT
datacenter_reference TEXT
tag_device_id VARCHAR(255)
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(11)
tag_coding_system TEXT
transmitted_id INT(11)
Indexes
otnunit_aat_animals
animal_project_reference TEXT
datacenter_reference TEXT
animal_reference_id VARCHAR(255)
vernacularname TEXT
scientificname TEXT
aphiaid INT(11)
tsn INT(11)
animal_origin TEXT
stock TEXT
length DOUBLE
length_type TEXT
weight DOUBLE
life_stage TEXT
age BIGINT(20)
sex TEXT
Indexes
otnunit_aat_detections
detection_project_reference TEXT
datacenter_reference TEXT
detection_id VARCHAR(255)
time TEXT
latitude DOUBLE
longitude DOUBLE
tracker_reference TEXT
detection_reference_id TEXT
detection_reference_type TEXT
transmitter_codespace TEXT
transmitter_id INT(11)
detection_serial_number INT(11)
sensor_data TEXT
sensor_data_units TEXT
deployment_id TEXT
position_data_source TEXT
Indexes
otnunit_aat_manmade_platform
platform_project_reference TEXT
datacenter_reference TEXT
platform_reference_id VARCHAR(255)
platform_type TEXT
platform_depth INT(11)
latitude DOUBLE
longitude DOUBLE
Indexes

Figure 1 – ERD diagram of OTN datasets before normalization

- **Figure 2** is the ERD generated using MySQL Workbench after normalization. There are a total of thirteen datasets.

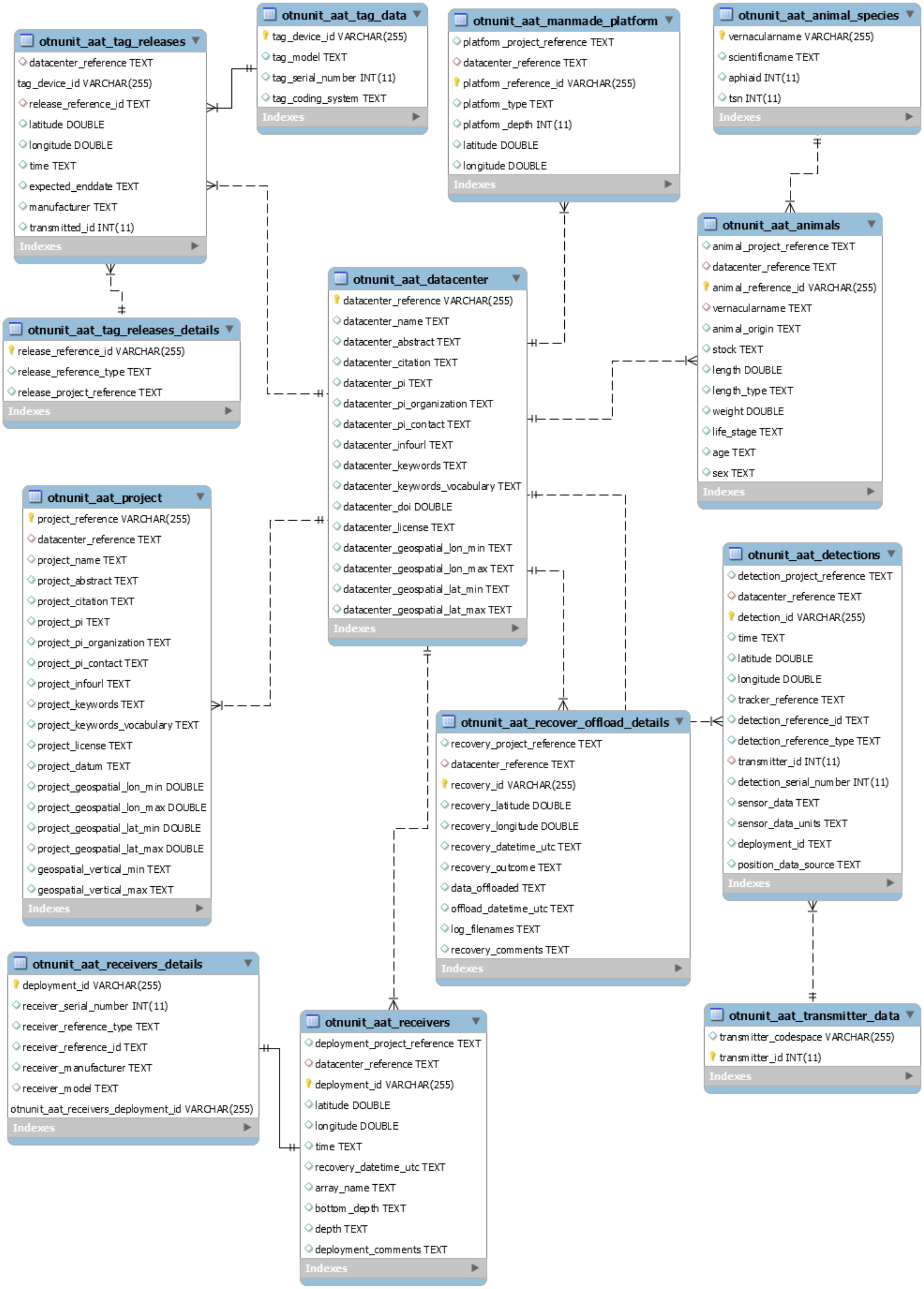
otnunit_aat_tag_releases
datacenter_reference TEXT
tag_device_id VARCHAR(255)
release_reference_id TEXT
latitude DOUBLE
longitude DOUBLE
time TEXT
expected_enddate TEXT
manufacturer TEXT
transmitted_id INT(11)
Indexes
otnunit_aat_tag_data
tag_device_id VARCHAR(255)
tag_model TEXT
tag_serial_number INT(11)
tag_coding_system TEXT
Indexes
otnunit_aat_manmade_platform
platform_project_reference TEXT
datacenter_reference TEXT
platform_reference_id VARCHAR(255)
platform_type TEXT
platform_depth INT(11)
latitude DOUBLE
longitude DOUBLE
Indexes
otnunit_aat_animal_species
vernacularname VARCHAR(255)
scientificname TEXT
aphiaid INT(11)
tsn INT(11)
Indexes
otnunit_aat_tag_releases_details
release_reference_id VARCHAR(255)
release_reference_type TEXT
release_project_reference TEXT
Indexes
otnunit_aat_datacenter
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_geospatial_lon_min TEXT
datacenter_geospatial_lon_max TEXT
datacenter_geospatial_lat_min TEXT
datacenter_geospatial_lat_max TEXT
Indexes
otnunit_aat_animals
animal_project_reference TEXT
datacenter_reference TEXT
animal_reference_id VARCHAR(255)
vernacularname TEXT
animal_origin TEXT
stock TEXT
length DOUBLE
length_type TEXT
weight DOUBLE
life_stage TEXT
age TEXT
sex TEXT
Indexes
otnunit_aat_project
project_reference VARCHAR(255)
datacenter_reference TEXT
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
geospatial_vertical_min TEXT
geospatial_vertical_max TEXT
Indexes
otnunit_aat_detections
detection_project_reference TEXT
datacenter_reference TEXT
detection_id VARCHAR(255)
time TEXT
latitude DOUBLE
longitude DOUBLE
tracker_reference TEXT
detection_reference_id TEXT
detection_reference_type TEXT
transmitter_id INT(11)
detection_serial_number INT(11)
sensor_data TEXT
sensor_data_units TEXT
deployment_id TEXT
position_data_source TEXT
Indexes
otnunit_aat_recover_offload_details
recovery_project_reference TEXT
datacenter_reference TEXT
recovery_id VARCHAR(255)
recovery_latitude DOUBLE
recovery_longitude DOUBLE
recovery_datetime_utc TEXT
recovery_outcome TEXT
data_offloaded TEXT
offload_datetime_utc TEXT
log_filenames TEXT
recovery_comments TEXT
Indexes
otnunit_aat_receivers_details
deployment_id VARCHAR(255)
receiver_serial_number INT(11)
receiver_reference_type TEXT
receiver_reference_id TEXT
receiver_manufacturer TEXT
receiver_model TEXT
otnunit_aat_receivers_deployment_id VARCHAR(255)
Indexes
otnunit_aat_receivers
deployment_project_reference TEXT
datacenter_reference TEXT
deployment_id VARCHAR(255)
latitude DOUBLE
longitude DOUBLE
time TEXT
recovery_datetime_utc TEXT
array_name TEXT
bottom_depth TEXT
depth TEXT
deployment_comments TEXT
Indexes
otnunit_aat_transmitter_data
transmitter_codespace VARCHAR(255)
transmitter_id INT(11)
Indexes

Figure 2 – ERD diagram of OTN datasets after normalization