# FORMAT OCEAN TRACKING DATA AND REPORT

Problem#2

Srikrishnan Sengottai Kasi
sr284605@dal.ca

Problem 2:

Datasets and Attributes:

Inferred from the website:

[Acoustic Receivers][1] OTN Receivers is considered a Data set with a unique id, location and information as its attributes.

Marine Scientists is a dataset with attributes like professional qualification, start Date, end date, designation id, address, email address and mobile.

Funding is a dataset to keep track of all the funds given for the growth of the project. Attributes will be like fund amount, fund provider, purpose of the fund, is fund is a loan.

Recorded data is a generic data set. This is classified based on the species. Attributes like species name, age, gender, origin, current location are tracked to serve the purpose of OTN.

OTN[1] council is a dataset. This gives importance to the historical values. The council has name, council id, member count. Alternatively, the scientist dataset can be generalized to employees and we can tag the council id if exists in that dataset as an attribute.

ISAC[1] members is also a dataset to store the various designations. It should store the designation id, designation name, voting status. This id can be reference in the employee's table.

Similar goes for SAC[1] and IDMC[1].

[Acoustic Tags][1] Taggers is a dataset to keep track of their current whereabouts. A unique id mapped to the taggers can be fixed so that when we track other species movements because of this tagger, then we can have an attribute like source received from tagger id.

Listener line is another dataset. It also carries the location, latitude, longitude and the server address to which the data has to be uploaded when an aquatic animal crosses the line.

Vemco Mobile Transceivers [VMT][1] is another dataset. The thing about that is it act as both receiver and sender. This is another key part in the Ocean Tracking Network. This consists of the location details, unique id, aquatic species id to which it is associated as an attribute.

Wave Gliders is another key entity in the OTN. This stores the weather data and oceanographic data as attributes. Its location is persisted as an attribute.

Observations from dataset:

The dataset is classified to three main segments, transmitter, receiver and deployer. Other than that, a separate dataset to the aquatic species is kept and the attributes associated with those objects are scientificname, vernacularname, origin, length, weight, gender and age.

In order to detect the aquatic species detectors are used. Each detection is recorded, and each detection are corresponded to unique datacentre. There are 4 datacentres across all the datasets and datacentre reference is almost used in all the datasets provided. The transmitter has its own properties like the id of the equipment and the type of encoding associated with it. The taggers also constitute of same attributes. It has the manufacture and the type of transmitter associated with it. Also, retrieval and recovery are another keen aspect in this process. Those are also considered as separate dataset and the time at which the recovery happened is logged. Along with that location data is stored for all the datasets.

Data cleaning:

otnunit_aat_animals_8dc3_4d15_c278.csv:

1) Taxon rank attribute is not recorded for any of the entries in the dataset and it doesn't derive any value in the dataset. So, it is removed completely.
2) Animal_guid seems to be a composite attribute. It can be subdivided to datacenter_reference + animal_project_reference + animal_reference_id.
3) animal_reference_id and animal_guid are the unique attribute across the 1309 records.
4) Since the guid can be constructed from the dataset we can remove that from the dataset to save some space. It will be easy to query the data with simple attributes than composite attribute. So, we can remove the animal_guid.
5) Even though it looks feasible to break the table based on the scientific name, the aphiaid and animal_project_reference is differing for one scientific name(CCSALMON and CCS) and the variation is based on the animal_origin(H and UNKNOWN) place. So, we cannot break the dataset. Animal_origin is a critical data from the receivers which is the main focus of the research study. So, it will be bad to assume it.
6) Based on the inference, when animal_project_reference, tsn, aphiaid, datacenter_reference and scientificname are same the origin remains the same. So, we can fill the missing origin based on that in the dataset. Even though there are some outliers like CCSALMON, we can take mean in the empty columns in few continuous records and it will not reduce the effectiveness.
7) Since most of the entries associated with the blue shark remains the same, we can assume that the stock is reported from the same region NWAtlantic for those records. The length_type and life_stage of most of the records is “FORK” and “JUVENILE” respectively. This can give us an idea that those species are moving in cluster across same listening lines and reported from the same region.
8) There are lot of unknown factors associated with GEFT animal project. So, the data will be corrupted if we make mean or remove the null values. We can however mark them as unknown on safer side to prevent false analysis in the future.
9) There are two values to denote unknown in the stock attribute. We will keep UNK to mark that. So, it is replaced accordingly in the dataset.
10) We cannot have NAN as a length value. This value is interpreted because of improper data type sent from the acoustic receivers and taggers. We can keep length as empty fields rather than NAN to avoid confusion while entering data to a relational table. It should be a number, but it will be absurd to take length as 0. Because no device can spot a species of length 0. Length Type is filled based on the length of the species. So, we cannot determine the length_type of those entries too. So, we can leave those blank.
11) The same logic goes for the weight and age attribute too. These attributes are critical to support the vision of the project. We will keep the values empty and will not add NOT_NULL constraint in the relational tables.
12) Unfortunately, age of almost 3658 records out of 3808 is unknown. So, it is better to remove that attribute as it will not support strongly in the analysis.
13) Out of the 3808 records, the value of sex is present for only 486 records. Even in that only 145 records have their sex determined. It is easy to mark all the blank one as unknown. But without determining sex in this dataset their population expansion and survival metrics become clueless. This will not contribute to their survival metrics in the changing climate condition in the waters. Due to very less information that can be derived from this attribute, removing this column from the dataset will not affect our analysis in the future.

14) Without proper information in the age, it will be difficult to point out its life stage. But life stage is present for considerably a greater number of species so we can fill empty columns with UNK as a value.

otnunit_aat_datacenter_attributes_8a94_cefd_f8a3.csv:

1) In this a lot of data are repeated. Vast information like abstract and license are repeated and stored in the records.
2) Additionally, we have empty columns which are critical but because of its empty values provides no value to the present dataset can be removed. The columns in that category are, date_modified, datacenter_distribution_statement, datacenter_date_modified, time_coverage_start, time_coverage_end.
3) NAN is not applicable to fields that takes number as input. So, they are made blank and optional while creating relational database. Those fields are datacenter_geospatial_lat_min, datacenter_geospatial_lat_max, datacenter_geospatial_lon_min. It will be bad to assume that information based on Mean or deviation because of the drastic variation in the location of the datacenter.
4) Fields like datacenter_pi_organization, datacenter_abstract, datacenter_pi, datacenter_pi_contact, datacenter_infourl, datacenter_keywords, datacenter_keywords_vocabulary, datacenter_doi, datacenter_license, remains the same for all four records. Those fields can be linked to the organization in a separate table. The organization can later be used to access those details.

otnunit_aat_detections_9062_5923_1394.csv

1) detection_guid is a composite attribute comprising of datacenter_reference + detection_id + detection_project_reference. So, select queries can use and operator to retrieve those records and the column can be deleted to eliminate repetition and save some space.
2) detection_reference_type stores only ANIMAL as field value. So that can be deleted from the spreadsheet and can be set as a default value while creating database. It is evident that OTN is tracking only ANIMAL in the ecosystem. So that column can be deleted as no predictions can be made from that field.
3) detection_transmittername is the combination of transmitter_codespace and transmitter_id. So, its again storing a composite attribute in the dataset which can be removed to reduce the redundancy and increase the storage.
4) receiver_log_id is full of NULL values and doesn't add any value to the collected dataset. So, we can remove the complete column from the dataset. This is a critical attribute to trace back in case of spotting future failures. But the system failed to record it. So removing the column for now from the provided dataset.
5) depth is another column that is filled with NaN fields. So, for the cleaning process we can remove that column from the dataset. But that was an important attribute to find the location of the species.
6) position_data_source entity carries only one value across the dataset. The value is "Receiver Metadata". The storage is required only when there are multiple values for a given field type. But there are more than 2 lakhs records carrying the same value. This reduces the fields importance in the dataset and can be removed. These values can be kept constant in the application layer itself. May be if there are different types of values in the future, we can add that in the dataset.

7) uncertainty_in_latitude, uncertainty_in_depth and uncertainty_in_longitude are another three disappointing fields in the dataset. Even from the field name we can understand that it is present to calculate any error that may have occurred from the readings sent by transmitter it doesn't add any significant value as an overall. In fact, it dilutes the precision of the information stored in the dataset. So, we can remove these two columns from the dataset to keep it clean. The same goes for depth_data_source, other_position_data, dataset_quality. They don't contribute anything for decision making in future for the recorded data. So, wiping them too.
8) detection_quality is another interesting attribute showcasing only one field value. "Found Receiver". But by the name of the field, it symbolises a quantitative measure than a string value. Because of the only one value recorded by that field in the dataset it loses its significance. In real time it plays a major role in deciding the authenticity of the data. Filling the empty values as "Not Available" will be more appropriate in this case.
9) sensor_data is empty for many rows. Based on the attribute name this data is recorded by the sensors deployed in the OTN. If no data is sent from the sensor it is safe to assume that as 0.0. As most of the fields are integers in this column.
10) Based on the detection_serial_number the latitude_degrees_north and longitude_degrees_east is detected. This shows the position of the detector which detects the species and send to the receiver. The table can be divided to store a mapping of detection_serial_number list and the latitude and longitude values associated with it.

otnunit_aat_manmade_platform_0735_7c9f_329c.csv:

1) platform_guid = datacenter_reference + platform_project_reference + platform_reference_id. So, this is a composite attribute and can be constructed later based on that logic. Hence, removing this column from the dataset to remove repetitive data.
2) platform_reference_id and platform_name is exactly one and the same. So, these records are considered duplicates in the dataset and platform_name can be removed. To avoid this the dataset should had different naming conventions for the platform_name.
3) NaN in platform_depth can be replaced with empty values. As clearly the depth can't be a string. So, we can have empty values for whatever the reading there is no depth information provided. Keeping it 0 will be wrong as it conveys a position.
4) This is a weak dataset as it doesn't have a single unique key to identify the record. So, bringing together latitude, longitude and platform_reference_id makes the values in the dataset look unique. Because of the null values present in the latitude and longitude makes it difficult to create an index. So, we can keep the platform_reference_id as the unique id and make the entity identifiable.

otnunit_aat_project_attributes_f29c_fb21_23a3.csv:

1) project_references, project_distribution_statement, project_date_modified and project_doi are completely null for all the records. They are project attributes associated with the project_reference. Moreover project_references and project_reference almost looks the same. So, removing it will not degrade the data. Same goes for geospatial_vertical_positive, time_coverage_start, time_coverage_end, project_linestring are also completely null and provides no use to the dataset.

2) project_infourl has few values that are improper. <NULL>, NA and some blank fields. So making it all same by filling the value as NA which represents Not Available for that field value.
3) project_pi_contact field is present for most of the columns. So, we can keep the empty values as Not Available(NA) for that column.
4) project_keywords_vocabulary, project_license, project_datum haven't varied based on any records. These types of static data can be associated separately in the application level. The license is pretty common across the OTN process and huge. The license column can be removed to reduce the redundancy and increase the storage.

otnunit_aat_receivers_c595_05f4_68b2.csv

1) expected_receiver_life is a field that is populated with NaN and blank values. This is an empty data in the dataset and provides no information about the receiver's lifespan. But this is an important attribute to be kept track. So, that we can stock our receivers accordingly in the future. Removing it in the spreadsheet to reduce the data.
2) deployed_by is another key attribute which is required to make a point of contact in the future. But the entire column is empty which makes the data insignificant and it can be removed from spreadsheet. But it can be used while creating a relational schema. This can be updated in the future.
3) deployment_guid is a composite attribute which comprises of datacenter_reference, deployment_project_reference and deployment_id. This creates repeated data, so it can be removed.
4) frequencies_monitored, receiver_coding_scheme are empty throughout the records and it is removed.
5) Receiver_serial_number values are tampered. It has few "?" and values like "can't read receiver number". Those values are replaced by NA for consistency.

otnunit_aat_recover_offload_details_4b23_f002_f89a.csv

1) recovery_guid = datacenter_reference + recovery_project_reference + recovery_id + deployment_id. It's a composite attribute and it can be derived from the above-mentioned simple attributes whenever needed. So, removing them to save some space.
2) clock_synchronized, recovered_by fields are completely null for records and makes it clueless to arrive at any decision based on the dataset. Thus, this can be added when sufficient information is recorded for it in the future.
3) There is a high resemblance in the deployment_id and recovery_id with few outliers. Even few deployment_id had some timestamp appended to it. Those are dirty data. From the list of deployment_id formats that is observed they seem to be more like a combination of the receiver_id and recovery_project_reference.

otnunit_aat_tag_releases_b793_03e7_a230.csv:

1) tag_programming_id, tag_frequency, and transmitter_type is completely null for all records. So, it can be removed from the dataset. In the case of taggers, they are attached with the transmitters. So, those attributes are important attributes and can be considered for the relational schema.
2) release_guid = datacenter_reference + release_project_reference + tag_device_id. Another composite attribute that can be removed for easy retrieval purposes in the future.

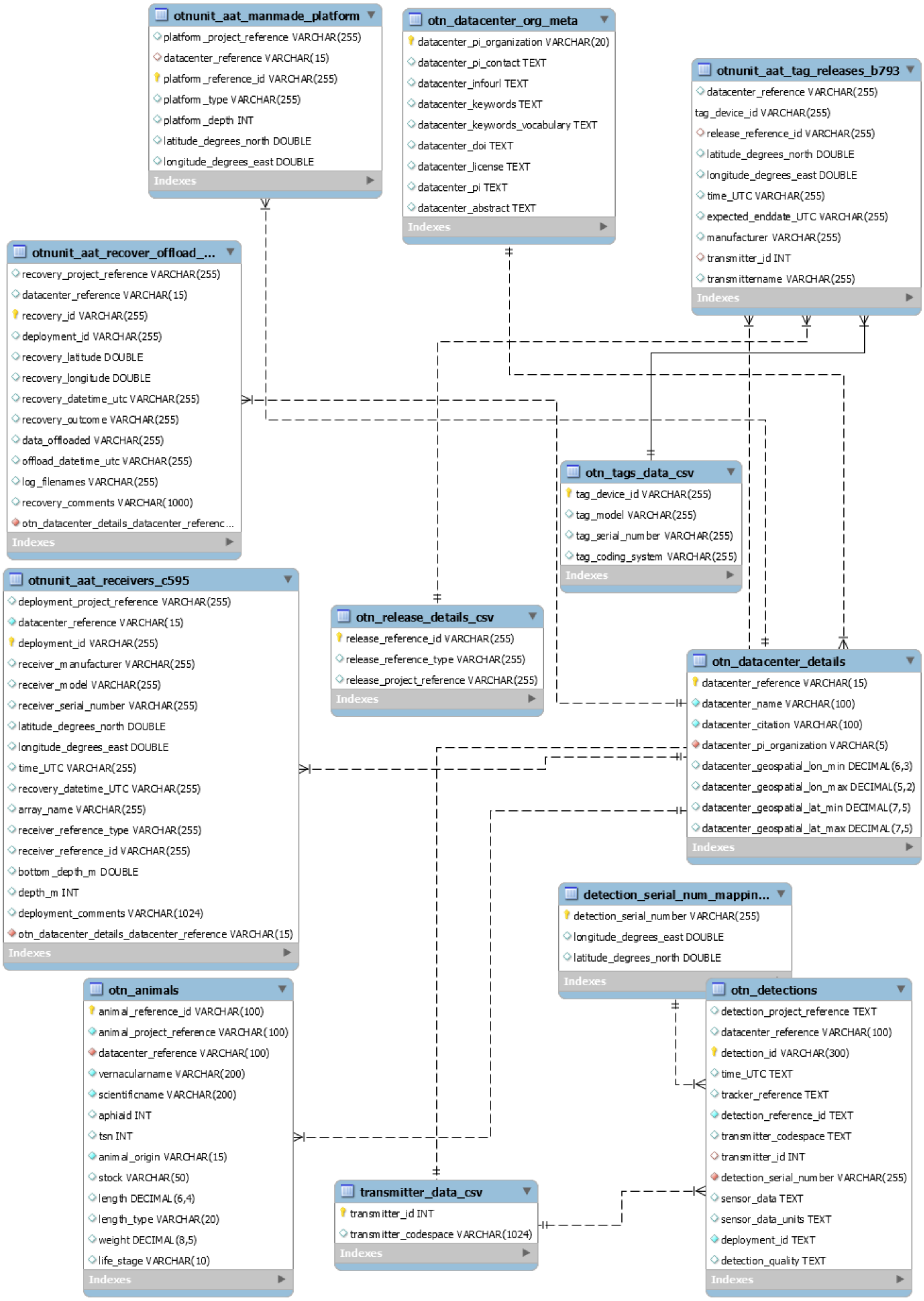
3) The table contains detailed information about the release and taggers. Hence, they are subdivided from the main dataset. We can keep the reference alone in the main dataset.
4) One table derived from it is release table. It will have release_reference_id as a primary id. It will carry release_reference_type and release_project_reference as other two fields. The release_reference_id alone can be stored in the main dataset.
5) Another part that can be divided is the tags. The tag_device_id can be the primary key of this table. It takes tag_model, tag_serial_number, tag_coding_system along with it.

In common the transmitter_id has been linked in two of the above datasets.

1) detectors
2) tagers

So, a separate decomposed dataset is made to map transmitter_id and the encoding pattern associated with it.

Reverse engineered ERD:


otnunit_aat_manmade_platform
platform_project_reference VARCHAR(255)
datacenter_reference VARCHAR(15)
platform_reference_id VARCHAR(255)
platform_type VARCHAR(255)
platform_depth INT
latitude_degrees_north DOUBLE
longitude_degrees_east DOUBLE
Indexes
otn_datacenter_org_meta
datacenter_pi_organization VARCHAR(20)
datacenter_pi_contact TEXT
datacenter_infourl TEXT
datacenter_keywords TEXT
datacenter_keywords_vocabulary TEXT
datacenter_doi TEXT
datacenter_license TEXT
datacenter_pi TEXT
datacenter_abstract TEXT
Indexes
otnunit_aat_tag_releases_b793
datacenter_reference VARCHAR(255)
tag_device_id VARCHAR(255)
release_reference_id VARCHAR(255)
latitude_degrees_north DOUBLE
longitude_degrees_east DOUBLE
time_UTC VARCHAR(255)
expected_enddate_UTC VARCHAR(255)
manufacturer VARCHAR(255)
transmitter_id INT
transmittername VARCHAR(255)
Indexes
otnunit_aat_recover_offload_...
recovery_project_reference VARCHAR(255)
datacenter_reference VARCHAR(15)
recovery_id VARCHAR(255)
deployment_id VARCHAR(255)
recovery_latitude DOUBLE
recovery_longitude DOUBLE
recovery_datetime_utc VARCHAR(255)
recovery_outcome VARCHAR(255)
data_offloaded VARCHAR(255)
offload_datetime_utc VARCHAR(255)
log_filenames VARCHAR(255)
recovery_comments VARCHAR(1000)
otn_datacenter_details_datacenter_referenc...
Indexes
otn_tags_data_csv
tag_device_id VARCHAR(255)
tag_model VARCHAR(255)
tag_serial_number VARCHAR(255)
tag_coding_system VARCHAR(255)
Indexes
otnunit_aat_receivers_c595
deployment_project_reference VARCHAR(255)
datacenter_reference VARCHAR(15)
deployment_id VARCHAR(255)
receiver_manufacturer VARCHAR(255)
receiver_model VARCHAR(255)
receiver_serial_number VARCHAR(255)
latitude_degrees_north DOUBLE
longitude_degrees_east DOUBLE
time_UTC VARCHAR(255)
recovery_datetime_UTC VARCHAR(255)
array_name VARCHAR(255)
receiver_reference_type VARCHAR(255)
receiver_reference_id VARCHAR(255)
bottom_depth_m DOUBLE
depth_m INT
deployment_comments VARCHAR(1024)
otn_datacenter_details_datacenter_reference VARCHAR(15)
Indexes
otn_release_details_csv
release_reference_id VARCHAR(255)
release_reference_type VARCHAR(255)
release_project_reference VARCHAR(255)
Indexes
otn_datacenter_details
datacenter_reference VARCHAR(15)
datacenter_name VARCHAR(100)
datacenter_citation VARCHAR(100)
datacenter_pi_organization VARCHAR(5)
datacenter_geospatial_lon_min DECIMAL(6,3)
datacenter_geospatial_lon_max DECIMAL(5,2)
datacenter_geospatial_lat_min DECIMAL(7,5)
datacenter_geospatial_lat_max DECIMAL(7,5)
Indexes
detection_serial_num_mappin...
detection_serial_number VARCHAR(255)
longitude_degrees_east DOUBLE
latitude_degrees_north DOUBLE
Indexes
otn_animals
animal_reference_id VARCHAR(100)
animal_project_reference VARCHAR(100)
datacenter_reference VARCHAR(100)
vernacularname VARCHAR(200)
scientificname VARCHAR(200)
aphiaid INT
tsn INT
animal_origin VARCHAR(15)
stock VARCHAR(50)
length DECIMAL(6,4)
length_type VARCHAR(20)
weight DECIMAL(8,5)
life_stage VARCHAR(10)
Indexes
otn_detections
detection_project_reference TEXT
datacenter_reference VARCHAR(100)
detection_id VARCHAR(300)
time_UTC TEXT
tracker_reference TEXT
detection_reference_id TEXT
transmitter_codespace TEXT
transmitter_id INT
detection_serial_number VARCHAR(255)
sensor_data TEXT
sensor_data_units TEXT
deployment_id TEXT
detection_quality TEXT
Indexes
transmitter_data_csv
transmitter_id INT
transmitter_codespace VARCHAR(1024)
Indexes

Citations and references:

[1] https://oceantrackingnetwork.org/about/#oceanmonitoring