CSCI 5408: Assignment 1

Problem 2

Document referred:
The following document http://oceantrackingnetwork.org/about/#oceanmonitoring is referred in identifying datasets and attributes.

Entities	Attributes	Reason for Selection
Employee	• Employee ID • First name • Last name • Date of Birth • Address	Considering the people that work at OTN on permanent or contractual basis. It is a strong entity.
Aquatic Species	• Species ID • Species Name • Scientific name • Vernacular name	Considering all the aquatic species that OTN is capturing data of. It is a strong entity.
Datacentre	• Datacentre ID • Datacentre Name • Citation • License • Location	Considering the datacentre as a strong entity as it stores and relays the information captured.
Acoustic Tag	• Tag ID • Tag Name • Tag manufacturer • Tag supplier	Considering the acoustic tag transmitters.
Acoustic Receiver	• Receiver ID • Receiver name • Receiver manufacturer • Receiver supplier	Considering the acoustic receivers.
VMT	• VMT ID • VMT Name • VMT manufacturer • VMT supplier	Considering the Vemco Mobile Transceivers.
Wave Glider	• WG ID • WG Name • WG manufacturer • WG supplier	Considering the wave glider as data transmitter to the satellite.
Slocum Glider	• SG ID • SG Name • SG manufacturer • SG supplier	Considering the slocum gliders as data transmitter to the satellite.

OTN Council	• Council Member ID • First name • Last name • Designation • Address	This is the management and support council that purview the ISAC, SAC and OTN Management Committee.
ISAC	• ISAC Member ID • First name • Last name • Designation • Address	This is a weak entity as it depends on the OTN Council.
SAC	• SAC Member ID • First name • Last name • Designation • Address	This is a weak entity as it depends on the OTN Council.
IDMC	• IDMC Member ID • First name • Last name • Designation • Address	This is a weak entity as it depends on the OTN Council.

The following datasets and attributes were discovered by me while going through the data dump provided.

Entities	**Attributes**
Animals	• animal_project_reference • datacenter_reference • animal_reference_id • animal_guid • vernacularname • scientificname • aphiaid • tsn • animal_origin • stock • length • length_type • weight • life_stage • age • sex
Datacenter_attributes	• datacenter_reference • datacenter_name • atacenter_abstract • datacenter_citation • datacenter_pi • datacenter_pi_organization

	• datacenter_pi_contact • datacenter_infourl • datacenter_keywords • datacenter_keywords_vocabulary • datacenter_doi, datacenter_license • datacenter_geospatial_lon_min • datacenter_geospatial_lon_max • datacenter_geospatial_lat_min • datacenter_geospatial_lat_max
Detections	• datacenter_reference,detection_id,detection_guid • time • latitude • longitude • tracker_ reference • detection_reference_id • detection_reference_type • transmitter_codespace • transmitter_id • detection_transmittername • detection_serial_number • deployment_id • depth • position_dat • source • uncertainty_in_latitude • uncertainty_in_longitude
Manmade_platform	• platform_project_preference • datacenter_reference • platform_reference_id • platform_guid • platform_type • platform_depth • platform_name • latitude • longitude
Project_attributes	• project_reference • datacenter_reference • project_name • project_abstract • project_citation • project_pi • project_pi_organization • project_pi_contact • project_infourl • project_keywords • project_keywords_vocabulary • project_references • project_doi • project_license • project_distribution_statement

	• project_date_modified • project_datum • project_geospatial_lon_min • project_geospatial_lon_max • project_geospatial_lat_min • project_geospatial_lat_max • project_linestring • geospatial_vertical_min • geospatial_vertical_max • geospatial_vertical_positive • time_coverage_start • time_coverage_end
Receivers	• deployment_project_reference • datacenter_reference • deployment_id • deployment_guid • receiver_manufacturer • receiver_model • frequencies_monitored • receiver_coding_scheme • receiver_serial_number • latitude longitude • time • recovery_datetime_utc • array_name • receiver_reference_type • receiver_reference_id • bottom_depth • depth • deployment_comments • deployed_by • expected_receiver_life
Recover_offload_details	• recovery_project_reference • datacenter_reference • recovery_id • deployment_id • recovery_guid • recovery_latitude • recovery_longitude • recovery_datetime_utc • recovery_outcome • data_offloaded • offload_datetime_utc • log_filenames • recovery_comments
Tag_ Releases	• 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 • tag_coding_system • tag_coding_system • transmitted_id • transmittername

1. **otnunit_aat_datacenter_attributes_8a94_cefd_f8a3.csv**

Operations performed:

- Removed columns **time_coverage_start** and **time_coverage_end** as they contained only one value against all the blank values in a row.
- Removed the 1st row (*after the column headings*) as it contained blank values against **time_coverage_start** and **time_coverage_end** columns.
- Under **datacenter_abstract** column, the data is in inconsistent format, so I formatted the other rows to make the data consistent.
- Under **datacenter_license** column, the data is in inconsistent format, so I formatted the other rows to make the data consistent.
- Removed **datacenter_distribution_statement** and **datacenter_date_modified** columns as they contained only blank values.
- Under the **datacenter_geospatial_lon_min, datacenter_geospatial_lon_max, datacenter_geospatial_lat_min, datacenter_geospatial_lat_max** columns the NaN values were replaced with values that don't come under the valid values range.

Analysis performed:

- Moved the **datacenter_citation** column and placed beside the **datacenter_name** column for better representation of the data.

2. **otnunit_aat_animals_8dc3_4d15_c278.csv**

Operations performed:

- Removed the 1st row (*after the column headings*) as it contained blank values.
- Removed column **taxonrank** as it has only blank values.
- Column **age, length** and **weight** have a lot of NaN so replacing NaN with -1. Doing this to protect other data present, instead of removing the column.
- For column **animal_origin** 12 values are blank for **scientific_name** Notorynchus cepedianus, so replacing those blank values with **W**.
- For column **stock** 22 values are blank for **scientific_name** Prionace glauca, so replacing those blank values with **NW Atlantic**.

- For column **stock** replacing all UNK values with **UNKNOWN**.
- For columns **life_stage**, **sex** and **length_type** replacing blank values with **UNKNOWN**.

Analysis performed:

- The **animal_guid** column is the combination of 3 columns **animal_project_reference, datacentre_reference** and **animal_reference_id**.
- Analysed the columns **stock**, **length**, **length_type** and **weight** values with respect to the Carcharhinus leucas under **scientific_name** and majority values are blank or NaN.
- Analysed the columns **length**, **length_type** and **weight** values with respect to the Galeocerdo cuvier under **scientific_name** and majority values are blank or NaN.

3. otnunit_aat_manmade_platform_0735_7c9f_329c.csv

Operations performed:

- Removed the 1st row (*after the column headings*) as it contained blank values for almost all and merged **degres_north** in latitude and **degrees_east** in longitude.
- Replacing NaN values in **platform_depth** with NA as it would be wrong to assume any random number.
- Replacing NaN values in **latitude_degrees_north** with -100 as it is out of range for the given values.
- Replacing NaN values in **longitude_degrees_east** with -200 as it is out of range for the given values.

4. otnunit_aat_project_attributes_f29c_fb21_23a3.csv

Operations performed:

- Removed the 1st row (*after the column headings*) as it contained blank values for almost all and merged the rows that had value with the heading.
- Replaced blank values with UNKNOWN in **project_abstract, project_citation, project_pi, project_pi_contact** column.
- Replaced blank and <NULL> values with NA in **project_infourl** column.
- Formatted **project_keywords** column.
- Deleting **project_references**, **project_doi, project_distribution_statement, project_date_modified, project_linestring, geospatial_vertical_positive, time_coverage_start** and **time_coverage_end** as they are completely blank.
- Replacing all the blank values in **geospatial_vertical_min** and **geospatial_vertical_max** with NA.

5. otnunit_aat_tag_releases_b793_03e7_a230.csv

Operations performed:

- Removed the 1st row (*after the column headings*) as it contained blank values for almost all and merged the rows that had value with the heading.
- Removing the **tag_frequency**, **transmitter_type** and **tag_programming_id** column as it is blank.

6. **otnunit_aat_receivers_c595_05f4_68b2.csv**

   Operations performed:
   - Removed the 1st row (*after the column headings*) as it contained blank values for almost all and merged the rows that had value with the heading.
   - Removing **frequencies_monitored, receiver_coding_scheme, deployed_by** and **expected_receiver_life** columns as they have all blank values.
   - Replacing all the blank and NaN values entries with NA.

7. **otnunit_aat_recover_offload_details_4b23_f002_f89a.csv**

   Operations performed:
   - Removed the 1st row (*after the column headings*) as it contained blank values for almost all and merged the rows that had value with the heading.
   - Removing **clock_synchronized** and **recovered_by**as they are empty columns.
   - Replaced all blank and NaN values with NA.

8. **otnunit_aat_detections_9062_5923_1394.csv**

   Operations performed:
   - Removed the 1st row (*after the column headings*) as it contained blank values for almost all and merged the rows that had value with the heading.
   - Removing **receiver_log_id, depth, uncertainty_in_latitude, uncertainty_in_longitude, depth_data_source, uncertainty_in_depth, other_position_data, dataset_quality** as they contain only NaN and blank values.
   - Replaced all blank values with NA.

Transformation done on data:

- There were many datasets that have alphanumeric values for the columns with numeric datatype. So, to avoid errors I changed the datatype of those columns to "**text**".
- I also faced an error in PK, having duplicate values so I changed that as well.

Normalization of the datasets:

- Normalization has been done for **receivers** table into **receivers** and **deployment** tables. Doing this will result in maintaining data integrity and simplifies the ERD.
- This operation is done as **deployment** table can be addressed as a separate entity.

The ERD diagram: