Name: Dixit Ghodadara
Banner ID: B00913652
Email: dx343670@dal.ca

COURSE: CSCSCI 5408 - Data Mgmt., Warehousing, Analytics

ASSIGNMENT – 1

Task.2 Ocean Research database discovery report

Datasets & attributes discovered

After analyzing the website http://oceantrackingnetwork.org/about/#oceanmonitoring these are discovered datasets and their attributes.

1. **Slocum Gliders** is dataset created by collection of wave information. Other attributes are id, location, model, manufacturer, and data received by various aquatic species.
2. **Predator swamping** is dataset created by analyzing behavior of salmon in high mortality landscape. Also, attached with journal with animal ecology. Contains attributes like predation id, type, location, area covered, landscape longitude and latitude.
3. **Marine Departments** is dataset with different attributes like dep. Id, description, manager id, joining date.
4. **Acoustic tags** is dataset contains attributes u id, code, attachment type, date of use, frequency and lifespan.
5. **Acoustic Receiver** is dataset with attributes receiver id, location, category, price, model, and data received by different aquatic species.
6. **Funding** is dataset with attributes like provider name, funding amount, approval date, given date, period, description, provider contact info.
7. **Data center** is dataset that collects information about the datacenters in OTN. Its attributes are center name, location, description, and number.
8. **OTN Council** is dataset for council. Contains these attributes. Member id, name, designation, email, organization and voting and nonvoting status of member.
9. **Wave Gliders** is dataset that gathers information on ocean and weather. Contains id, location updates, model, manufacturer, data received by different aquatic species.
10. **Marine scientist** is dataset for scientists' information. Contains attributes like unique id, name, dob, address, email, joining date, leaving date, department id and supervisor id.

Data Pre-processing & Normalization

1) **otnunit_aat_animals_8dc3_4d15_c278.csv**
- Removed 1st null row and taxorank column.
- Normalization of age is dependent on animal_project_reference created.
- Removed age to remove redundancy and NaN data without losing any data.
- Normalized life_stage
- Animal_project_reference can be removed as it is from animal_ref_id. This may cause over-normalization as it takes too much processing resources to fetch.

2) **otnunit_aat_datacenter_attributes_8a94_cefd_f8a3.csv**
- Removed null columns datacenter_distribution_statement, datacenter_date_modifies, time_coverage and time_coverage_end.
- Removed NaN values from columns datacenter_geospatial_lon_min.
- Removed blank space from tesco.
- Added same values in columns datacenter_abstract, datacenter_citation, datacenter_pi, datacenter_keywords, datacenter_keyword_vocab, datacenter_doi and datacenter_license

3) **otnunit_aat_detections_9062_5923_1394.csv**
- First row specifies that latitude and longitude are in degrees north and east respectively. These are not data but data types, so record deleted.
- Composite attribute is also deleted.
- The column sensor_date had empty values. But it was crucial factor to measure data so, I retained the column, but I changed the blanks to NULL values.
- Empty rows updated with NA.
- The column receiver_log_id was empty, thus removed.
- There were only NaN values in depth. Hence, it deleted completely.

4) **otnunit_aat_manmade_platform_0735_7c9f_329c.csv**
- The first row again specifies that lat. and long. Values in degree. These values are not data but data type. Hence, this record is deleted.
- The column platform_id is composite attribute that is combination of column. Hence, this column is removed.
- Platform depth had many NaN values. Depth cannot have non-numeric values. Hence this data replaced with NULL.
- The platoform_ref_id and platform_name columns are the same. There is no need for both. I removed platform_name from the dataset.

5) **otnunit_aat_project_attributes_f29c_fb21_23a3**
- Again, first row specifies that latitude and longitude values. These are not the data but data types. Hence, it deleted.
- Empty columns like project_reference, project_doi and time_coverage and so on. Deleted from the dataset.

- Project_abstract had blank cells, filled with NA.
- Null values in different columns replaced with NA for consistency.
- The blanks were replaced with NULL as it is for geo_vertical_min and geo_vertical_max.

6) **otnunit_aat_receivers_c595_05f4_68b2.csv**
- Again, first row filled with latitude and longitude in degrees. Bottom_depth and depth are in meters. These values are not the data but data types for mentioned columns. Hence, this was deleted.
- The column deployment_guide is composite attribute that is a combination of columns.
- Other three columns were completely empty. Hence, I removed.
- Deployment_commnets column was filled with empty. They filled with UNK flag for consistency.
- The column bottom_depth consists of NaN values which were replaced by NULL.

7) **otnunit_aat_recover_offload_details_4b23_f002_f89a.csv**
- The first row is empty; hence this row is deleted.
- Composite attributes column is deleted. So, it can be derived from other 3 columns.
- The columns clock_synchronized and reovered_by are completely empty. Hence, I deleted both from dataset.
- There are some percentage values missing from columns, I filled empty cells.

8) **otnunit_aat_tag_releases_b793_03e7_a230.csv**
- Again, first row specifies that latitude and longitude values. These are not the data but data types. Hence, it deleted.
- The column release_guide is composite attribute that is a combination of columns datacenter_reference, release_project_reference, tag_device_id. Hence, this column removed.
- The column transmittername is composite attribute so, deleted it. So, It can be derived from other 2 columns.
- The empty columns are removed from the dataset. These columns can add back when we have sufficient data to fill.

Normalization process:

1) **animals**
- This dataset was already in 1NF.
- The columns vernacularname, scientificname, tsn and aphiaid were added in separated dataset called “animal_spieces”. Vernacularname act as foreign key in animals table.

2) **detections**
- This was also in 1NF.
- The columns tra_codespan and id were added in separate dataset called transmitter_data. Id act as foreign key in detections table.

3) **tag_releases**
- This was already in 1NF.
- The columns release_ref_id, release_ref_type and other 2 columns were added in separate dataset called tag_release_details. reference_id act as foreign key in tag releases.
- The column tag_devices_id, tag_model, tag_serial_number and tag_coding_system was added in separated dataset called tag_data. tag_device_id acts as a foreign key for tag_release.

4) **receivers**
- This was already in 1NF.
- The columns dep_id, receiver_serial_number, ref_id, model, and other fields added to separated dataset called receiver_details. This is on one-to-one relation with receivers.

ERD generated using MySQL workbench without normalization. (8 base tables)

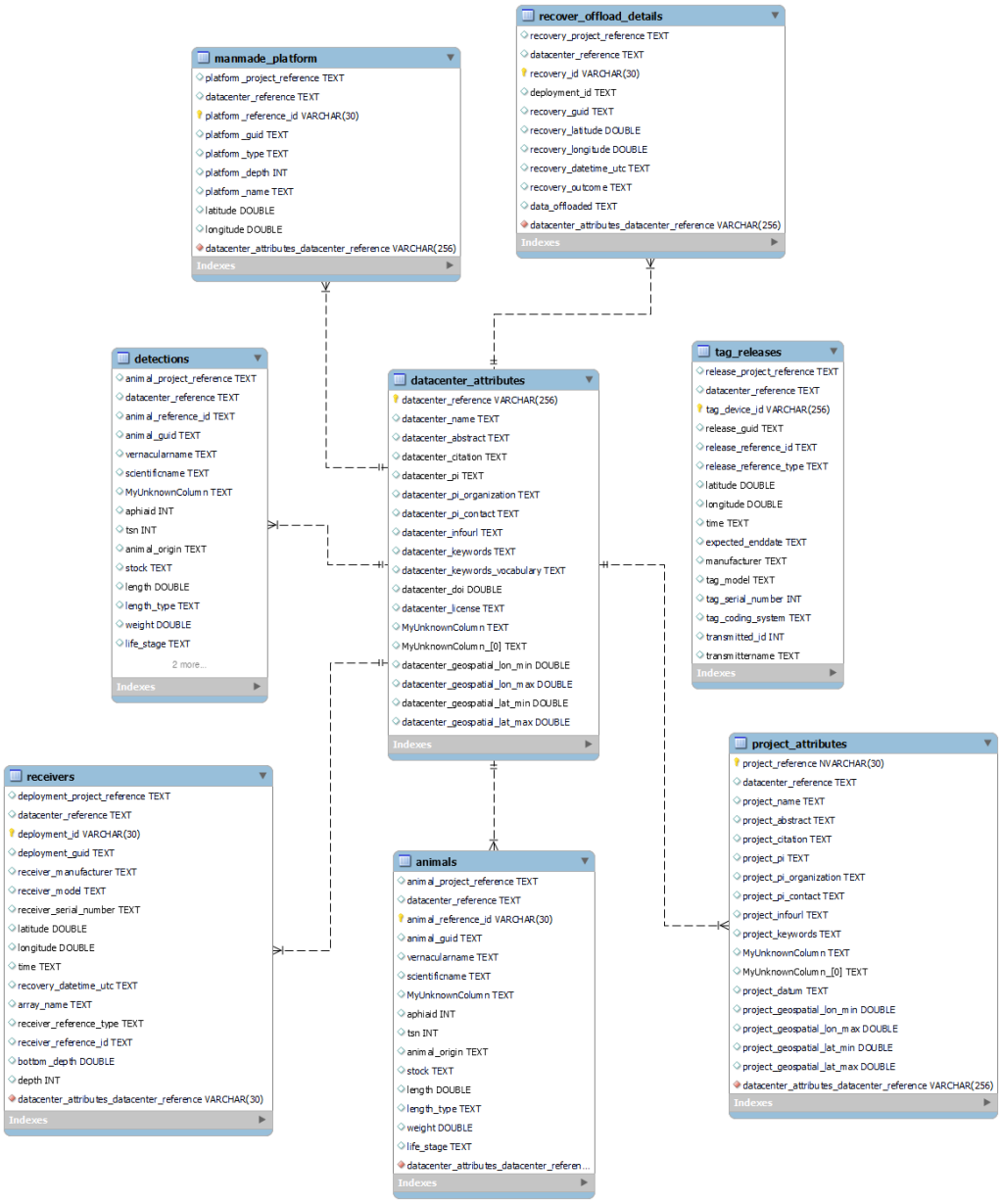

Figure 1 Denormalized ERD

ERD generated using MySQL workbench with normalization. (13 tables)

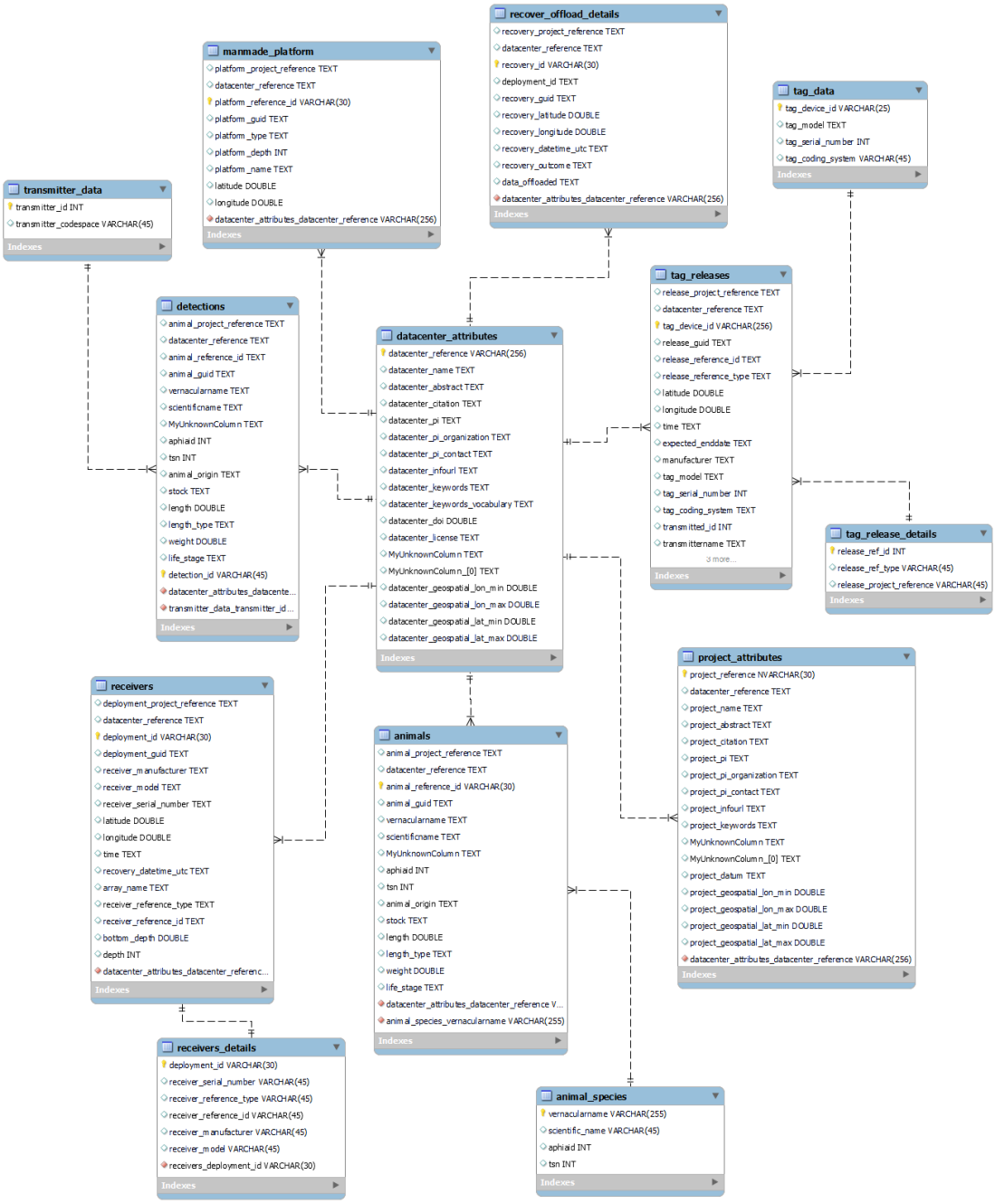

Figure 2 Normalized ERD

Reference(s):

[1] **Oracle Corporation, "MySQL Workbench", September 2005. [Online]. Available: MySQL Workbench, https://www.mysql.com/products/workbench [Accessed: 13 September, 2022].**

[2] **"Diagrams.net - free flowchart maker and diagrams online," *Flowchart Maker & Online Diagram Software*. [Online]. Available: https://www.draw.io/index.html. [Accessed: 21-Sep-2022]**