CSCI 5408

Assignment 1

Problem 3

Datasets & Attributes

The following objects are possible Datasets and Attributes that can be conceptualized from the article available at: https://oceantrackingnetwork.org/about/#oceanmonitoring

1. Equipment:

Attributes:

a) Equipment_Id

b) Equipment Type

c) Date Procured

d) Manufacturer

e) Total Cost

2. Marine Animals:

Attributes:

a) Species_Id

b) Species Name

c) Scientific Name

d) Taxonomy_Family

e) Taxonomy_Class

f) Phylum

g) Tag

h) Gender

i) Age

3. Employee:

Attributes:

a) Employee_Id

b) Employee_Name

c) Date of Birth

d) Gender

e) Department_Id

f) Designation

g) Date_Recruited

h) Contact Number

i) Email Id

j) Location

4. Acoustic Tracking Station:

Attributes:

a) Station_Id

b) Station Name

c) Latitude

d) Longitude

e) Taxonomy_Class

f) Transmitter Model

g) Receiver Model

h) Mobile Receiver Model

i) Transmitter Weight

j) Software

5. Funds

Attributes:

a) Contributor

b) Cause

c) Sum

d) Utilized Amount

e) Balance

6. Committee

Attributes:

f) Committee_Id

g) Committee Name

h) Number of members

i) Committee Description

j) Date_Formed

7. Acoustic Tag

Attributes:

a) Tag_Id

b) Transmitter Model

c) Range

d) Transmitter Weight

e) Manufacturer

f) Frequency

8. Acoustic Receivers

Attributes:

a) Receiver_Id

b) Receiver Model

c) Receiver Weight

d) Ship_Id

9. Mobile Transceivers

Attributes:

a) Transceiver_Id

b) Transmitter Model

c) Mobile Receiver Model

d) Transmitter Weight

e) Receiver Weight

10. Wave Gliders

Attributes:

a) Glider_Id

b) Latitude

c) Longitude

11. Ship

Attributes:

a) Ship_Id

b) Current Latitude

c) Current Longitude

d) Company

12. Signal

Attributes:

a) Signal_Id

b) Latitude

c) Longitude

d) Depth

e) Animal_Id

Dataset Transformation

This section serves as documentation for the steps involved in cleaning & transforming the datasets associated with Problem 3.

1. otnunit_aat_animals_8dc3_4d15_c278

I) Delete the entire column "taxonrank" since it does not contain values for any rows.

II) Replace Null values in the column "stock" with the string "Null".

III) Replace NaN (Not a Number) values in the column "length" with the string "Null" to maintain consistency across the file.

IV) Replace Null values in the column "length_type" with the string "Null".

V) Replace NaN (Not a Number) values in the column "weight" with the string "Null" to maintain consistency across the file.

VI) Replace Null values in the column "life_stage" with the string "Null".

VII) Replace NaN (Not a Number) values in the column “age” with the string “Null” to maintain consistency across the file.

VIII) Replace Null values in the Column “sex” with ‘U’ under the assumption that ‘U’ refers to unidentified.

IX) Delete the second row of the file since the entire row is empty/blank.

X) **Normalize the file into two separate files to eliminate partial dependencies**:

i. Animals_Taxonomy(animal_project_reference,datacenter_reference, animal_reference_id, **animal_guid,** vernacularname, scientificname, aphiaid, tsn, animal_origin and stock. (**Fields in Bold are foreign keys**)

ii. Animals_Physical_Attributes (**animal_guid**, length, length_type, weight, life_stage, age & sex)

XI) Rearrange the columns such that the subsequent columns lead to a GUID column. Example: Rearrange [animal_project_reference, datacenter_reference, animal_reference_id and animal_guid]

2. otnunit_aat_datacenter_attributes_8a94_cefd_f8a3

I) Remove special characters from the rows under the column “datacenter_abstract”.

II) Remove special characters from the rows under the column “datacenter_license

III) Delete the columns datacenter_distribution_statement & datacenter_date_modified since all the rows under these columns are blank.

IV) Replace NaN with Null under the columns: datacenter_geospatial_lon_min, datacenter_geospatial_lon_max, datacenter_geospatial_lat_min & datacenter_geospatial_lat_max.

V) Delete the columns time_coverage_start & time_coverage_end since they’re empty.

VI) Shift the column “datacenter_pi_organization” before the first column and the column “datacenter_citation” before the third column, since these columns are similar and convey the same information.

3. otnunit_aat_detections_9062_5923_1394

I) Replace Null values in the Column “sensor_data” with the string “Null”.

II)Replace special characters and blanks with the value “metres” under the column “sensor_data_units”.

III)Replace Null values in the Column “detection_quality” with the string “Null”.

IV)Replace NaN with Null under the column “depth” with Null.

V)Delete the columns – “uncertainty_in_latitude” & “uncertainty_in_longitude” since all the rows contain NaN as values.

VI) Delete the columns - receiver_log_id , depth_data_source, uncertainty_in_depth, other_position_data & dataset_quality since all rows under these columns are blank.

VII) Delete the second row since it offers no business value.

VIII) Normalize the tables, to prevent partial dependency, into:

i. otnunit_aat_detections_9062_5923_1394 with columns (**detection_guid**, time, latitude, longitude, tracker_reference detection_reference_id, detection_reference_type, detection_reference_type, transmitter_codespace, transmitter_id and detection_transmittername). (Fields in Bold are foreign keys)

ii. Detection_Transmitter_Information with the rest of the columns.

4. otnunit_aat_manmade_platform_0735_7c9f_329c

I) Replace NaN with the string "Null" in the column – "platform_depth".

5. otnunit_aat_project_attributes_f29c_fb21_23a3

I) Rearrange columns - project_reference, project_name and project_infourl since they are related.

II) Rearrange columns project_keywords_vocabulary & project_datum to the beginning of the file since they are common for all rows.

III) Delete the columns - project_references, project_doi, project_distribution_statement, project_date_modified, geospatial_vertical_positive, time_coverage_start and time_coverage_end since they contain blank values for all rows.

IV) Replace nulls with the value 0 for the column geospatial_vertical_min and with the string "Null" for the column geospatial_vertical_max.

6. otnunit_aat_receivers_c595_05f4_68b2

I) Normalize the file into two tables(fields in Bold are foreign keys)::

i. Receiver_information with columns (datacenter_reference, deployment_id, **deployment_guid**, receiver_manufacturer, receiver_model, receiver_serial_number,, receiver_reference_type and receiver_reference_id).

ii. otnunit_aat_receivers_c595_05f4_68b2 with the rest of the columns.

II) Delete the columns - frequencies_monitored, receiver_coding_scheme, deployed_by and expected_receiver_life since these columns are completely blank.

III) Replace NaN values with Null for the columns - bottom_depth and depth.

7. otnunit_aat_recover_offload_details_4b23_f002_f89a

I) Normalize the file into two tables (fields in Bold are foreign keys):

i. Recovery_Information with columns (recovery_project_reference, datacenter_reference, recovery_id, deployment_id and **recovery_guid**)

ii. otnunit_aat_recover_offload_details_4b23_f002_f89a with the rest of the columns

II) Rearrange the columns in the file "Recovery_Information" such that they contiguously form the **recovery_guid.**

III) Handle Null values in the columns - recovery_datetime_utc, recovery_outcome, data_offloaded, offload_datetime_utc, log_filenames and recovery_comments.

IV) Delete the columns - clock_synchronized and recovered_by since these columns are completely null and offer no business value.

8. otnunit_aat_tag_releases_b793_03e7_a230

I) Normalize the file into three tables (fields in Bold are foreign keys):

i. Tag_Information with columns (release_project_reference, datacenter_reference, tag_device_id, tag_model, tag_serial_number and **release_guid**)

ii. Transmitter_Information with columns tag_coding_system, transmitted_id and **transmittername**

iii. otnunit_aat_tag_releases_b793_03e7_a230 with the rest of the columns

II) Delete the columns - tag_frequency, transmitter_type and tag_programming_id since they are completely Null.

References

1. CSCI 5408 Assignment 1 Handout
2. CSCI 5408 Lecture Slides

Declaration

I Benny Daniel Tharigopala, declare that in assignment 1 of CSCI 5408 course, data scrapping is not done programmatically or using any online or offline tools. However, the webpages or the domain mentioned in this document are visited manually, and some useful information is gathered for education purpose only. Information, such as email, personal contact numbers, or names of people are not extracted. The course instructor or the Faculty of Computer Science cannot be held responsible for any misuse of the extracted data.