Dugong Aerial Survey Database Design Document

Steven Vandervalk
eResearch Centre, James Cook University
06.05.2015

The following is a technical document aimed at database administrators or designers. It outlines the structure of the database and details the assumptions, tradeoffs and rationale for the design.

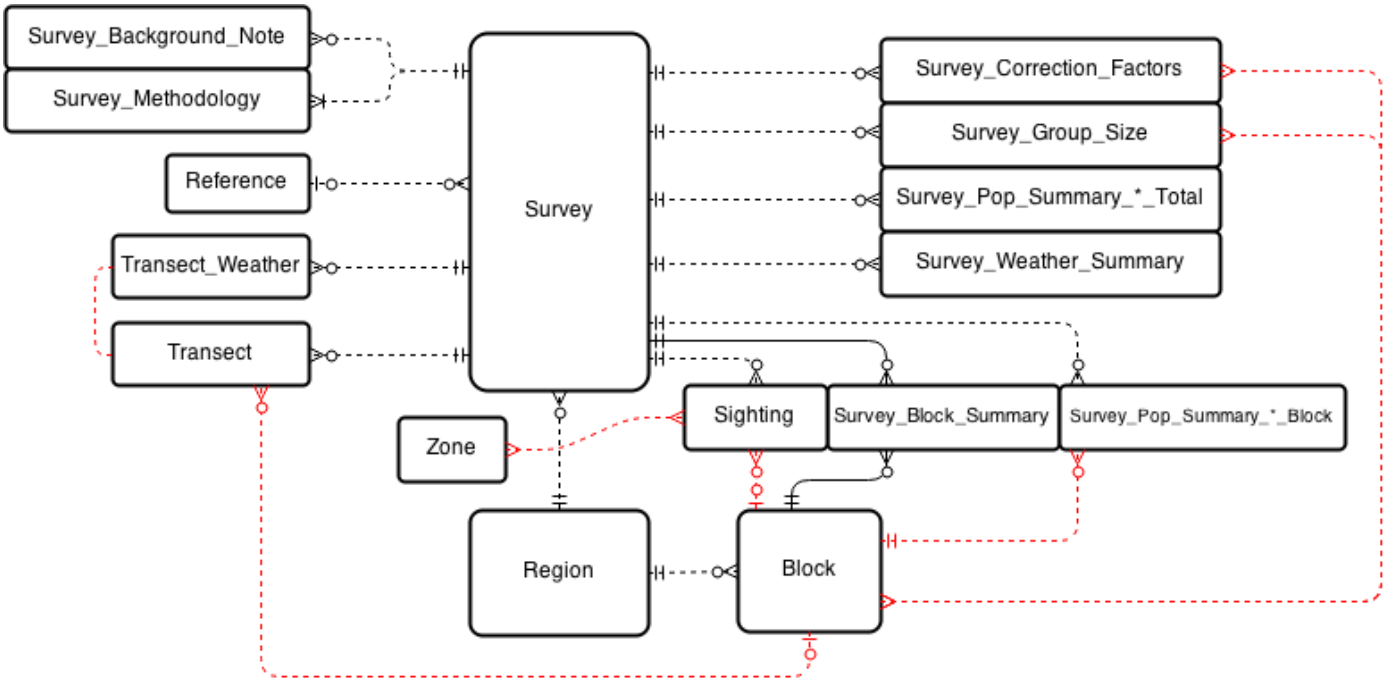
The database is currently hosted at https://dugongs.tropicaldatahub.org/ and maintained by the eResearch Centre, James Cook University.

The database was iteratively designed in collaboration with marine biologists that were the original data custodians and end users. Due to the nature of the data and a strong need for end-user ease of use, the database was primarily normalised[1] to third normal form[2] (3NF), with strict referential integrity not upheld in a few instances. Additionally, example queries were provided with the end user documentation to allow users not familiar with SQL to use both as-is and adapt to suit their own queries.

The document is organised as follows : section 1 contains a broad overview of database structure and relationships using a standard entity relationship diagram[3] with Crow's foot notation[4]. Section 2 splits the tables into three categories which shows how they were viewed logically. Section 3 contains the workflow to create the database from primary data to MySQL. Finally, section 4 lists each individual table and explains any assumptions and rationales it contains, the fields and their properties.

Section 1

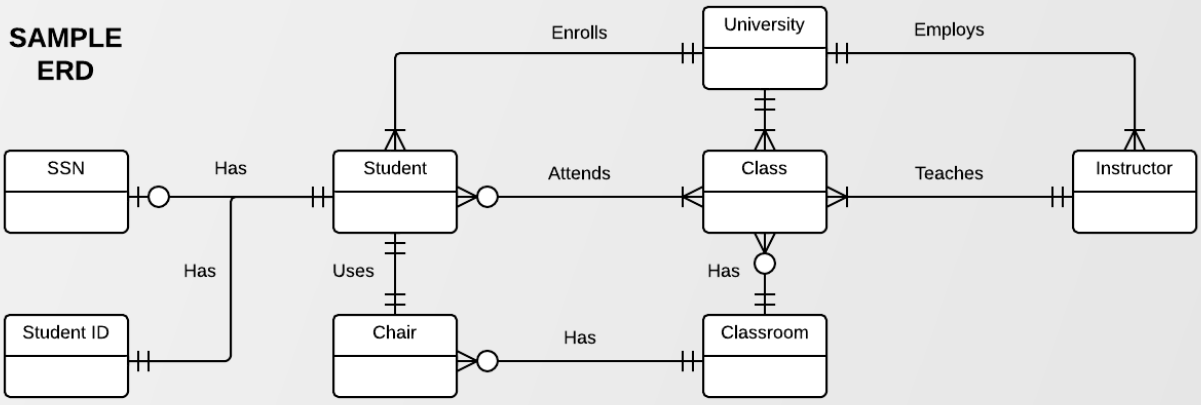
Entity Relationship Diagram



[1] https://en.wikipedia.org/wiki/Database_normalization
[2] https://en.wikipedia.org/wiki/Third_normal_form
[3] https://en.wikipedia.org/wiki/Entity%E2%80%93relationship_model
[4] https://en.wikipedia.org/wiki/Entity%E2%80%93relationship_model#Crow.27s_foot_notation

ERD "Crow's Foot" Relationship Symbols [Quick Reference]

Created by Vivek M. Chawla | @VivekMChawla | April 7, 2013



Notation	Meaning	Example
	Relationship	Student — Enrolls — University
	One	Student — Has — Student ID Number
	Many	Student — Attends — Class
	One and ONLY One	Student — Uses — Chair
	Zero or One	Student — Has — Social Security Number
	One or Many	Instructor — Teaches — Class
	Zero or Many	Classroom — Has — Chair

Table of tables

Meta	Observation	Summary
Block	Sighting	Sighting_Calves_Summary
Reference	Transect_Flight	Survey_Block_Summary
Region	Transect_Weather	Survey_Correction_Factors
Survey		Survey_Group_Size
Survey_Background_Note		Survey_Pop_Summary_Marsh_Sinclair_Block
Survey_Count_Type		Survey_Pop_Summary_Marsh_Sinclair_Total
Zone		Survey_Pop_Summary_Pollock_Block
		Survey_Pop_Summary_Pollock_Total
		Survey_Weather_Summary

Section 2

Data import workflow

1. Archival material to spreadsheet (MS Excel)
2. Spreadsheet to standard spreadsheets (Google Sheets)
3. Standard spreadsheet to CSV files
4. Edit CSV file to have a trailing newline
5. Import into MySQL table

Section 3

Tables

Block

Block records the associations between blocks and their regions. It may seem unnecessary, given that the block_id incorporates the reg_id presently, but it should be recorded semantically in the database, and not just in the syntax/structure of the strings.

Fields

block_id (PK)

reg_id (FK, ref Region)

Reference

References to papers. ID field, citation and URL.

Fields
ref_id (PK)
ref_citation
ref_url

Region

ID, region name, and URL for a ZIP file of shapefiles for the region.

Fields
reg_id (PK)
reg_name
shapefiles_url

Sighting

Information about dugong and herd sightings. **sighting_type** identifies the sighting as a dugong (or multiple dugongs) or a herd.

Fields
sighting_id (PK)
surv_id (FK, ref Survey)
sighting_type
tf_num_in_survey
trans_id
block_id
time
observer
number_in_group
number_at_surf
number_break_surf_or_deep
break_or_deep
zones
calves
latitude
longitude
turbidity
comments

Sighting_Calves_Summary

Summary information about calves sightings in surveys.

Fields
cs_id (PK)
surv_id (FK ref Survey)
perc_calves

num_calves
num_total
comments

Survey

ID, region and reference for the Survey. ID currently formed as Region-Year-Month(-Letter). Month is usually the month in which the survey began. In some cases, the month is dropped as the survey took place at different times of the year. In one case, a letter is appended for survey in the same region and month, but the data is to be considered separate (MB-2000-12-A vs MB-2000-12-B). Dates and methodology are stored separately in **Survey_Methodology**, because there may be separate parts to the survey.

Fields
surv_id (PK)
reg_id (FK, ref Region)
surv_start
surv_end
flight_height
transect_width
observers

Survey_Background_Note

Background info about surveys is recorded as notes in this table.

Fields
note_id (PK)
surv_id (FK, ref Survey)
note_text

Survey_Block_Summary

Area information about a surveyed block.

Fields
surv_id (PK, FK ref Survey)
block_id (PK, FK ref Block)
area
sampling_intensity
perp_zone_length
total_possible_transects
comments

Survey_Correction_Factors

For some surveys, there are different correction factors for different parts of the region. **cf_id** is necessary to identify each row of the table. **section** lists the particular blocks/transects that the correction factors apply too, if it's not the entire surveyed region.

perc = perception
avail = availability
cf = correct factor
cv = coefficient of variation
p/f = port/starboard
f/r/b = front/rear/both

Fields

cf_id (PK)
surv_id (FK, ref Survey)
perc_port_cf
perc_port_cv
perc_starboard
perc_starboard_cv
perc_num_pf
perc_num_pr
perc_num_pb
perc_num_sf
perc_num_sr
perc_num_sb
avail_cf
avail_cv
dug_groups_lessthan_ten_surface
dug_groups_lessthan_ten_under
dug_groups_lessthan_ten_total
section
comments

Survey_Group_Size

Group size info. **section** may be a single block, multiple blocks, or “total”. **gss_id** is needed to uniquely distinguish each row.

Fields

gss_id (PK)
surv_id (FK, ref Survey)
section
group_size
gsize_coef_var
gsize_std_err

Survey_Methodology

Records start and end dates, and methodology info for an entire survey, or part of a survey. **meth_id** is needed to uniquely distinguish each row.

Fields

meth_id (PK)
surv_id (FK ref Survey)
surv_start
surv_end

flight_height
transect_width
observers
comments

Survey_Pop_Summary_Marsh_Sinclair_Block
Survey_Pop_Summary_Marsh_Sinclair_Total
Survey_Pop_Summary_Pollock_Block
Survey_Pop_Summary_Pollock_Total

Fields
surv_id (PK, FK ref Survey)
[block_id (PK, NOT FK because contains some non-blocks - QLDtotal, NTtotal)] not in “Total” tables
nhat
nhat_std_error
comments

Survey_Weather_Summary

For some regions, weather summaries are broken up into different sections for some surveys, so **ws_id** is needed to uniquely identify each row.

Fields
ws_id (PK)
surv_id (FK ref Survey)
section
max_wind_speed
min_cloud_cover
max_cloud_cover
min_cloud_height
beaufort_mean_of_modes
beaufort_min
beaufort_max
glare_north_mean_of_modes
glare_north_min
glare_north_max
glare_south_mean_of_modes
glare_south_min
glare_south_max
glare_overall_mean_of_modes
min_visibility
comments

Transect

trans_row_id is needed to uniquely identify each row in the table. **tf_num_in_survey** is the transect’s number within its shapefile. **trans_id** is an ID for the *location* where the transect is being done.

Fields

trans_row_id (PK)
surv_id (FK, ref Survey)
block_id (NOT FK, as some transects do not belong to a single block, see GoC-1991-12 for example)
tf_num_in_survey
trans_id
date
start_lat
start_long
end_lat
end_long
height
area
length
start_time
end_time
direction
comment

Transect_Weather

Weather data recorded during transects.

Fields

weather_id (PK)
surv_id (FK, ref Survey)
tf_num_in_survey
trans_id
beaufort_min
beaufort_max
beaufort_mode
glare_north_min
glare_north_max
glare_north_mode
glare_south_min
glare_south_max
glare_south_mode
comment

Zone

Records zone identifiers (V, H, M, L, O, I, A) against the description (“very high”, “high”, etc.)

Fields

zone_id (PK)
description