First TUFMAN Training Workshop (TTW-1)
3-7 December 2012
SPC, Noumea, New Caledonia

Session 11.1

Data Quality

Data Entry Checks

INTRODUCTION

TUFMAN has a lot of checks that occur as data are being entered, to try and catch data entry errors before they are saved. While the data entry checks have been made as stringent as possible, it is important to note that these will not catch all data entry errors so it is still important to run data quality checks after the data have been entered.

What types of data entry errors can occur?

There are probably too many to mention, and if we tried then there would certainly be some entry errors that have been missed. However here is an example of the types of errors that can commonly occur:

- 'Transposition' error - where numbers or characters are entered in the wrong order, the key here is that the same characters or numbers are entered.
    - e.g. 112 entered as 121
- 'Transcription' error – any case of what is written on the form being entered differently into the database.
    - 112 entered as 122
    - 1-jan-2012 as 1-jan 2021
    - 101 entered as 1001
    - Etc
- Missing data – data have not been entered when they should have, such as catch information for a set or catch for a particular species
- Duplicate data – where the same thing has been entered twice
- Recording errors – this is where the data on the form itself is actually wrong, so it has been written down incorrectly by the recorder
- Formatting errors – the most common here is with dates since 10/11/12 with US settings is different to 10/11/12 with English/Australian settings

Trapping data entry errors

The main principles of 'trapping' (i.e. identifying) data entry errors are:

- Restrict to a range of values where possible, for example:
    - Catch in numbers will never be negative, and should not exceed a maximum value
    - Dates may not be allowable in the future (e.g. logsheet sets)
    - Set dates should be within the departure and return dates of a trip
    - Average weight of a single fish could be limited with a minimum and maximum
- Only allow certain characters to be entered if possible – i.e. if a number is being entered then you should only be allowed to enter a number, not letters.

- Remove ambiguity by using formatting – for example dates are formatted to '10-Nov-2012' format, so if you enter as 10/11/12 you can see that the date has been interpreted correctly
- Use format 'masks' if possible - for example the entry of positions, these are restricted to strict entry formats for the latitude and longitude.
- Use other data to verify what has been entered, for example the previous set, other dates (e.g. first logdate must be between departure and return dates)

A lot of errors can be caught by applying these rules, but they will not catch all errors. For example if 112 is entered instead of 122 then it would be impossible to detect this error if 112 is a valid number. Data entry checks usually catch extreme values when it comes to numbers.

The data entry checks in TUFMAN

This section describes some of the typical data entry checks in TUFMAN. It doesn't list all since there are too many to list and a lot of them are probably obvious. Other types of data are checked in a similar way to those below.

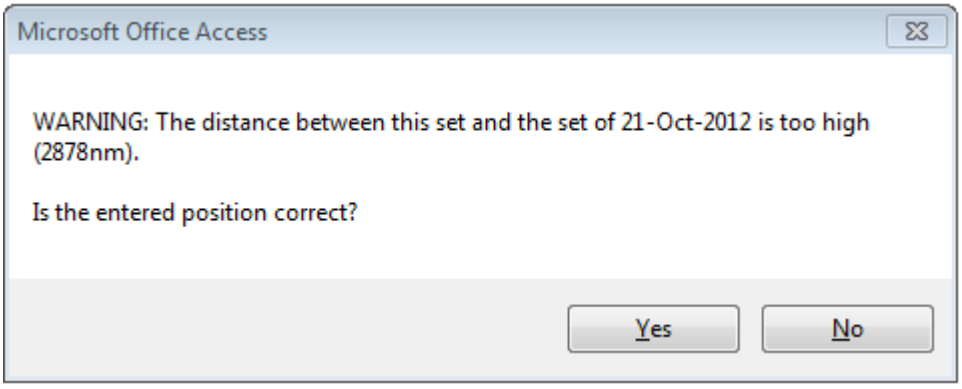
Logsheet data checks

Trips:

- Dates in proper order (e.g. depart_date < return_date etc.)
- Dates don't overlap with another trip for the vessel
- Departure date to first logsheet date check, should be less than a certain value
- Last log date to return date check, should be less than a certain value
- Trip Length check
- Dates cannot be in the future

Sets:

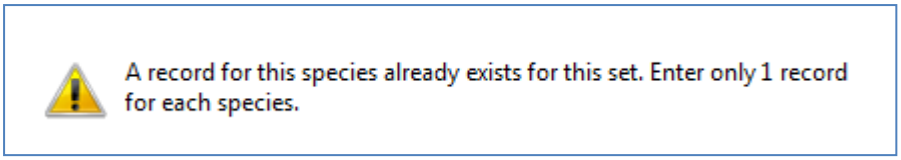
- Set dates within trip dates
- No overlap with another set within a certain time period (gear dependent)
- Set distance limits check (distance between 2 sets should not exceed a certain speed that it would take to travel that far within the time between the 2 sets). Will give a warning if it is too high but allow the data to be saved:



- Number of hooks limit check
- Hooks between float limits check
- Time <24 hours, minutes < 60 (e.g. 25:65 not allowed)
- Latitude > 90 not allowed, minutes must be < 60 (e.g. 9100N, 7961W not allowed)
- Longitude > 180 not allowed, minutes must be < 60 (e.g. 18100W, 17961W not allowed)

Catch:

- Species range checks – Numbers and weights of fish by species less than a defined limit for that species
- average weights, catch by weights checks – must be within pre-defined limits for the species
- Only 1 row per species per set



Licensing

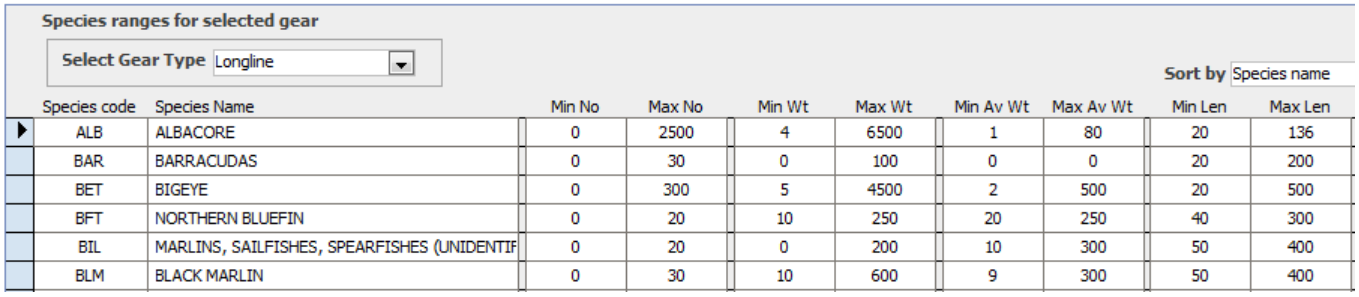
- License start date must be within the dates of the agreement reporting period
- License dates mustn't overlap with another license for the same vessel
- License number should be unique

Vessel details:

- IRCS must not contain non-alphanumeric characters
- Vessel name should be unique
- Registration number should be unique

The Species Ranges table in TUFMAN

The species range table contains information that helps do range checking on numbers and weights entered by species. The table is by gear since there can be large differences between fish caught by different gear types. The species range table is accessed via the 'Admin->Reference tables->Species ranges' menu item:

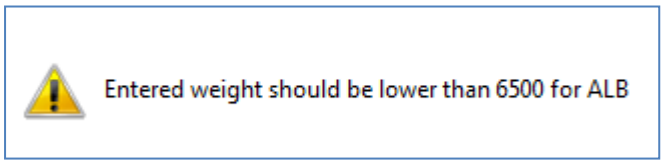


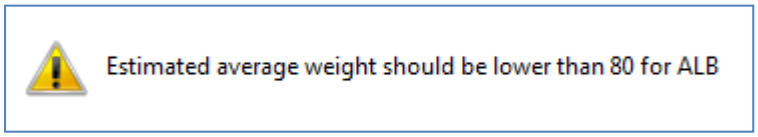
Species code	Species Name	Min No	Max No	Min Wt	Max Wt	Min Av Wt	Max Av Wt	Min Len	Max Len
ALB	ALBACORE	0	2500	4	6500	1	80	20	136
BAR	BARRACUDAS	0	30	0	100	0	0	20	200
BET	BIGEYE	0	300	5	4500	2	500	20	500
BFT	NORTHERN BLUEFIN	0	20	10	250	20	250	40	300
BIL	MARLINS, SAILFISHES, SPEARFISHES (UNIDENTIF	0	20	0	200	10	300	50	400
BLM	BLACK MARLIN	0	30	10	600	9	300	50	400

- **Min No** –the minimum number of fish allowed in a set
- **Max No** - the maximum number of fish allowed in a set

- **Min Wt** –the minimum weight of fish (kg) allowed in a set
- **Max Wt** - the maximum weight of fish (kg) allowed in a set
- **Min Av Wt** –the minimum average weight of fish (kg) allowed in a set. Average weight is calculated by dividing catch in weight by catch in numbers
- **Max Av Wt** - the maximum average weight of fish (kg) allowed in a set
- **Min Len** – The minimum length (cm) for port sampling
- **Max Len** – The maximum length (cm) for port sampling

These species range tables are use by TUFMAN during data entry of logsheet and port sampling data. If a species catch is entered then TUFMAN will compare the values to the range in this table, and give an error message if the value is outside of the range. If the species is not in this table then it will accept all values. The table is 'local' to your country, which means that you can change it in your country if you find that some of the limits are too low. For example the maximum catch per set for ALB in the screenshot above is 6500kg. If you had a super-fleet that could catch up to 8000kg per set then you could change the value from 6500 to 8000.





The 'limits' table in TUFMAN

In TUFMAN there is another table which contains information used for data entry checking, which is called the 'limits' table. This is a table designed to be flexible in holding different range values and allowing easy addition of others. Here is an example of some data from this table:

gear	Variable	Descriptor	Min_Error	Min_Warning	Max_Warning	Max_Error
L	hooks_n	Number of hooks	150			6000
L	Hooks_Btn_floats	Hooks between floats	4	0	0	50
L	Trip_Length	Trip length	0	0	60	360
L	Depart_to_Log	Number of days between departure and first log date	0	0	3	40
L	Log_to_Return	Number of days between the last log date and return to port	0	0	3	40
L	Settime	Number of hours between sets	0	15	0	0
L	SetDistance	Distance between consecutive sets	0	0	0	250
S	Trip_Length	Trip length	0	0	365	0
S	Depart_to_Log	Number of days between departure and first log date	0	0	3	0
S	Log_to_Return	Number of days between the last log date and return to port	0	0	3	0
S	Settime	Number of hours between sets	0	0	0	0
S	SetDistance	Distance between consecutive sets	0	0	0	360
S	Return_to_Sample	Number of days between return date to sample date	0	0	7	0
S	Hours_Samp	Number of hours taken to sample the fish	0	0	0	149
S	WellSamp_Mt	Weight of sampled wells in metric tonnes	0	0	0	200
P	Trip_Length	Trip length	0	0	60	360
P	Depart_to_Log	Number of days between departure and first log date	0	0	7	160
P	Log_to_Return	Number of days between the last log date and return to port	0	0	7	160
P	SetDistance	Distance between consecutive sets	0	0	0	250
L	LastLog_to_Unload	Number of days between last logdate to unloading date	0	0	7	14
L	Return_to_Sample	Number of days between return date to sample date	0	0	7	0
L	Hours_Samp	Number of hours taken to sample the fish	0	0	0	24

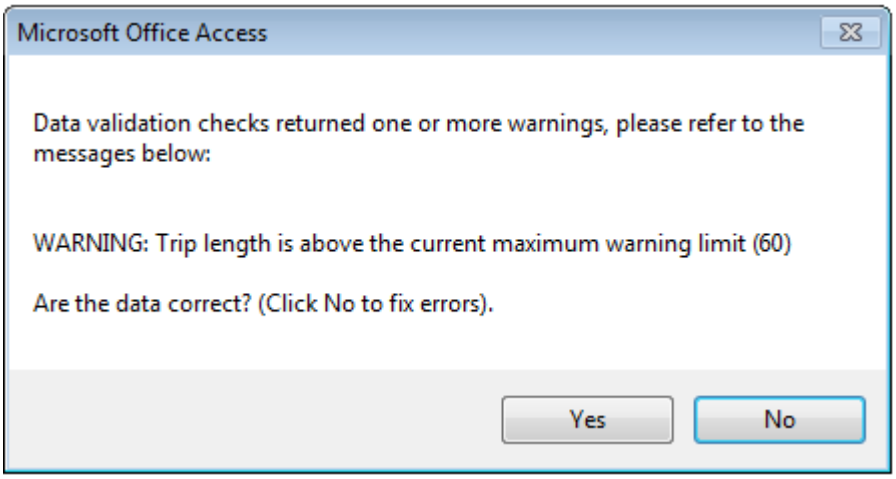
The fields of the table are:

- **Gear –** each row is for a specific gear
- **Variable** – a short name for the type of check the entry is for

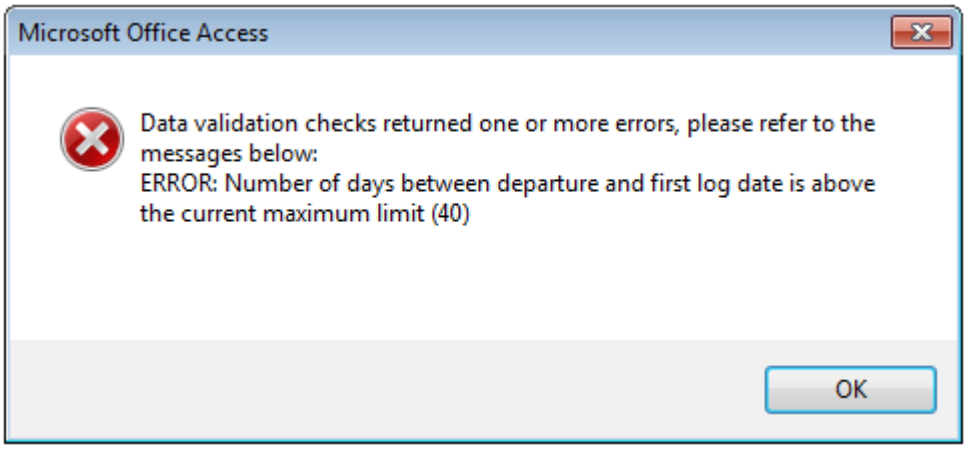
- **Descriptor** – description of the check
- **Min_Error** – defines the lowest allowable value. A value lower than min_error will cause an error and the data will not be allowed to be saved.
- **Min_Warning** – A value lower than min_warning will cause a warning to be raised, but the user is allowed to save once they double-check the entered value.
- **Max_Warning** – A value larger than max_warning will cause a warning to be raised, but the user is allowed to save once they double-check the entered value.
- **Max_Error** – A value greater than max_error will cause an error and the data will not be allowed to be saved.
- **Comments** (not shown) – some extra comments describing what the check is for

When data are entered, TUFMAN will check this table and compare the entered value to the min and max warning and error values, and will raise any appropriate messages. Usually TUFMAN will check several values at once and report on them all at the same time. For example if a longline record is entered with:

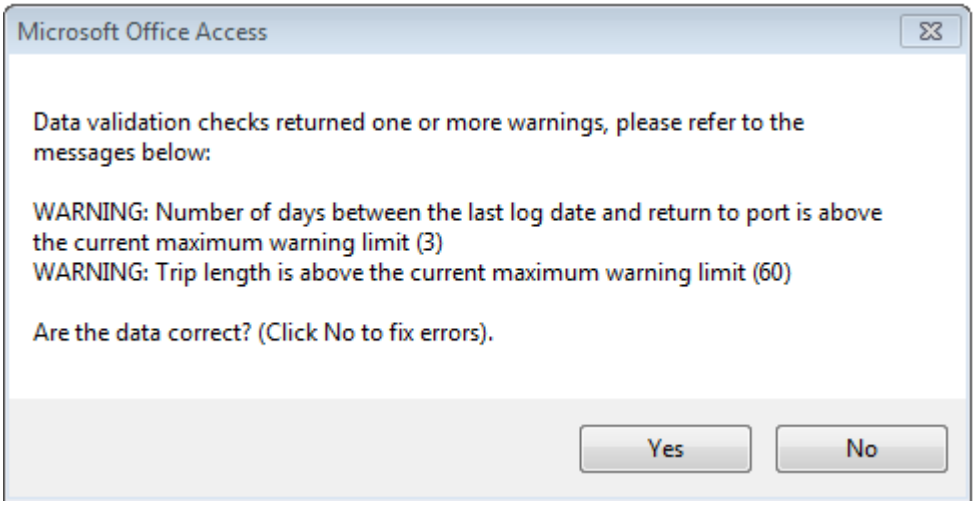
- Trip length > 60 but less than 360 (warning – can click 'yes' to keep what was entered, No to change the entered value if it was wrong)



- First logdate > 40 days after date of departure (error so must change the values)



- A combination of 2 warnings:



In the table, the error and warning values have been calculated by looking at a lot of existing data:

- Error values – these have been calculated by finding the values that almost never occur, e.g. there are almost no longline trips than last longer than 360 days (but there are exceptions)
- Warning values – these are values that can occur but are rare, so say 95% of values should be within the warning range of values

These values are set at SPC and are the same for all users of TUFMAN. If you find that data cannot be entered because TUFMAN says the values are too high or too low, but the value is correct, then SPC can change this table or create an exception to allow the data to be entered.

SPC will be improving this table in the near future by allowing the values to be set for a particular flag and country, since there can be large differences in values from one flag to another, and one country to another. The lower the maximum values are, the more data entry errors are trapped, so it is best to have these values as 'tight' as possible.

Logsheet Check Totals

Logsheet check totals are a different type of data entry check that is applied when all sets and catch for the trip have been entered. The idea is to add up all catch for the trip by species, and verify this value by entering the totals for the trip recorded on the logsheet. This is a very good method of data quality checking and is better than using the species range table, since it will detect a difference as low as 1 as long as the totals have been calculated correctly.

SOUTH PACIFIC REGIONAL LONGLINE LOGSHEET

PAGE ____ OF ____

10

Clutch

Clutch

NZ

NZ 1031

FISHING PERMIT OR LICENCE NUMBER(S): 004

NAME OF AGENT IN PORT OF UNLOADING: NFP

- ALL DATES AND TIMES MUST BE UTC / GMT
- ALL WEIGHTS MUST BE KILOGRAMS

PORT OF DEPARTURE: Alofi

PORT OF UNLOADING: Alofi

PRIMARY TARGET SPECIES: Yello Fin

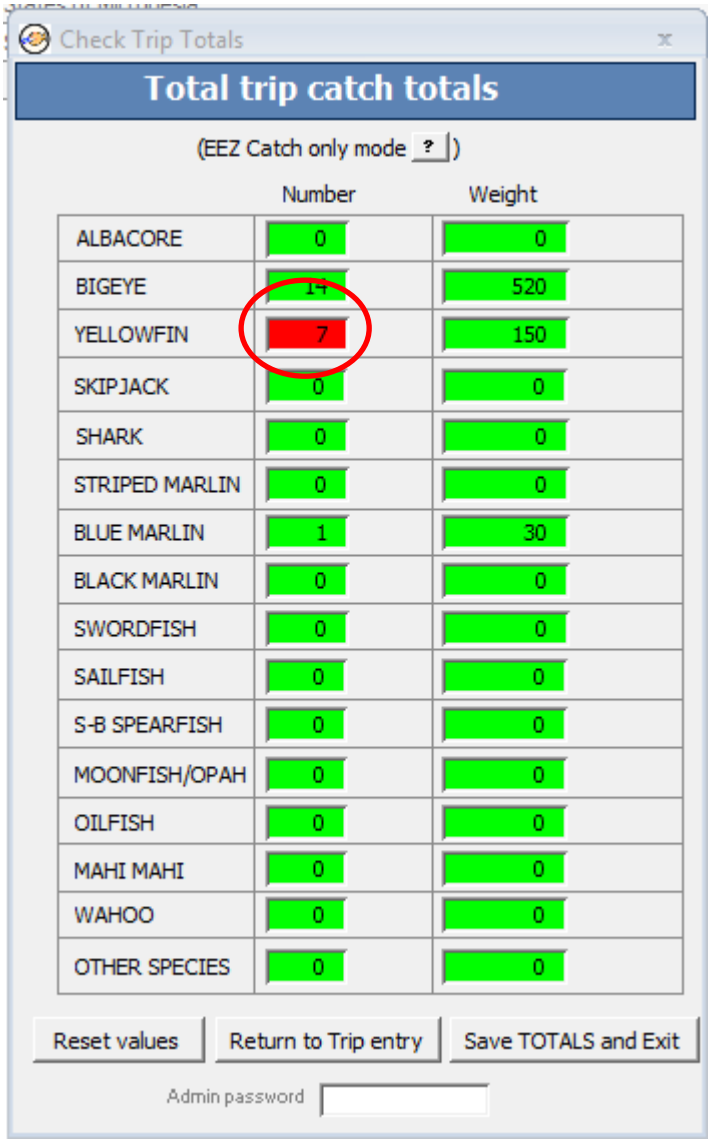
YEAR: 2005

DATE AND TIME OF DEPARTURE: 12/07/05 0630

DATE AND TIME OF ARRIVAL IN PORT: 16/07/05 0530

NUMBER OF HOOKS BETWEEN FLOATS: 35

01.00 UTC OR SET POSITION							SET START TIME	NUMBER OF HOOKS	ALBACORE			BIGEYE			YELLOWFIN			WAHOO			SHARK		STRIPED MARLIN		BLUE MARLIN		BLACK MARLIN		SWORDFISH		OTHER SPECIES		
DATE	LATITUDE DDMM	S	LONGITUDE DDDMM	E/W					No RET	KG RET	No DISC	No RET	KG RET	No DISC	No RET	KG RET	No DISC	No RET	KG RET	No DISC	No RET	No DISC	No RET	KG RET	No RET	KG RET	No RET	KG RET	No RET	KG RET	NAME	No RET	KG RET
07 12	2	18 54	S	170 13	W		0730	1000	21	300		—	—		4	70		—	—		—		—	—	1	40	—	—	—	—	SKJ/Mahi Esc	13	78
07 13	2	19 00	S	170 13	W		0730	1000	10	150		—	—		2	35		—	—		—		—	—	—	—	—	—	—	—	SKJ/Mahi Short/Esc	5	40
07 14	2	19 00	S	170 11	W		0800	1000	7	100		3	90		3	70		1	6		2		—	—	—	—	—	—	1	7	SKJ/Esc/Mahi	8	50
07 15	2	18 55	S	170 11	W		0800	1000	14	200		5	130		3	90		2	12		1		—	—	—	—	—	—	—	—	Mahi/SKJ Short	5	25
ACTIVITY CODES							PAGE TOTAL		52	750		8	220		12	265		3	18		3				1	40					SKJ/Mahi	31	193
							TRIP TOTAL		52	750		8	220		12	265		3	18		3				1	40					Esc/Short	31	193



Vessel Trip Track plot

This is a feature of TUFMAN that has been around for a while, but has recently been moved and can now be launched from the logsheet detail form, and also from the logsheet summary list. It allows for a visual check of the vessel track for the logsheet, which can help identify badly entered or badly recorded positions, like the position highlighted below: