RE: Priority fields for vessel census database

dangleloo@gmail.com <dangleloo@gmail.com>

Mon 12/14/2020 12:38 PM

To: So, Dane (FAOKH) <Dane.So@fao.org>; Brown, David (FAOKH) <David.Brown@fao.org>
Cc: 'SamAth Leng' <lengsamathy@gmail.com>

Dear Dane,

Thanks for the update. The fields Roitana mentioned are the key fields for using the vessel register for licensing. To get those field 'clean', i.e. getting the values verified and correct for all records, may be impossible without going back to the owner/fisher. Even the original questionnaire may not help, because there was insufficient QA/QC during data collection.

My priority list was more inclusive, because we need to clean and allow access to most data in the database, not just the data that will be used for licensing. However, if FIA wants to focus on these few core fields first, that is not a problem, because we have very limited time.

The flow chart for the system looks fine. I like that it can allow for public access first (without registration), before asking for login to access actual databases. We can streamline that later.

I **agree** that you should focus on getting a working PHP system going that allows to access the SQL database, whatever data sets FIA decides to use.
I **suggest** you assess how many records are 'clean' for the core fields as identified by Roitana, so that we have an idea about what still needs to be done, unless that is not possible.

- Did you discuss the targets under the NPCI with Roitana?
- Did you discuss which records to do first? As we discussed before, it makes sense to start with the larger vessels (>18 meters and work your way down, as that is the order in which the licensing will proceed.
- I already assumed that IDNo would need to be the national ID number, but some people may not have answered that, so then you need to at least get the name, address and contact details correct.
- The actual setup of the hosting/network is required when the vessel census database will go on-line, but as we discussed, we will keep it off-line for the moment, as sharing the data with FIAC and WMPT depends on a decision by the DG

You write "migrate the current database to current system", that is not entirely clear. Does that mean you will setup an SQL database based on Roitana's Excel data and link that to the PHP GUI you are developing? If so, that is fine.

I don't see how you can clean the data, without write access, unless you document all proposed changes, make a separate database. Perhaps if FIA needs to do the data cleaning themselves, you should train some staff how to do it themselves. The whole point of FAO support was to assist in cleaning the database first and then to develop the web-interface to access it. It sounds if FIA now wants to have a full interface first, then will work on cleaning the data and finally deciding on the fields to include, depending on different end-users.

We need to be clearer on our recommendations in view of the targets, e.g. prepare a report with an assessment of the main issues and recommended approach to fix them. We need to document the whole process, including decisions and agreements, so it is clear what was done and why. If you don't have time, perhaps I can assist.

Best regards,
Theo

From: So, Dane (FAOKH) <Dane.So@fao.org>
Sent: 14 December 2020 11:23
To: dangleloo@gmail.com; Brown, David (FAOKH) <David.Brown@fao.org>
Cc: 'SamAth Leng' <lengsamathy@gmail.com>
Subject: Re: Priority fields for vessel census database

Dear Theo,

after consultation with Mr. Roitana and Sam Ath,

now, I have no right to editing rowData, just leave data as it is (which i really spent so much time on developing script programming, check consistency, just edit). Now, moving to the next step.
- He suggest few key important such as 1. Census No which needs to clean, 2. National ID of Owner, 3. Length, 4. Power Engin and 5. Fishing gear

(Noted- រឿងField សំខាន់ដែលយើងត្រូវសំអាតមុន មាន៖ ១. លេខជំរឿន ២.ព័ត៌មានម្ចាស់ទូក (អត្តសញ្ញាណប័ន....) ៣.ប្រវែងទូក ៤.កំលាំងម៉ាស៊ីន ៥.ឧបករណ៍នេសាទ។

WHERE Item2 IDNO in the SERVER (some only). good thing, there is potential reformate Census ID in the server, so, I am still trying to do.

I don't have enough time and just focus on these key first and also develop GUI (PHP and MySQL) which i have work sofare to make it function first as attached navigation prototype here.

Clean those priority, develop form and migrate data into database and bring it to SERVER of FIA. I am try to finalize these early next week and will rely on NETWORK of FIA and authorization of Mr. Roitana.

As i have mention in another email about admin access to network, it does not rely on me. It is completely other services (VPN, IP Router/Getaways to assign port to work etc). With talk with IT Mr. Sambo, current control by two company,

Public IP EZECOME https://www.ezecom.com.kh/

NETWORK Company manage router/ port (official first bid but letter they change to alternative with former staff to new company, that another need to learn more), http://www.trust-groups.com/

to save my time and finalize things, please give me focus this week to be able to migrate the current database to current system.

attached yours my progress on the prototype.

ONCE Database complete with new function, they can be editing what ever available or update new.

Regards

Dane

From: dangleloo@gmail.com <dangleloo@gmail.com>
Sent: Friday, December 11, 2020 12:35 PM
To: So, Dane (FAOKH) <Dane.So@fao.org>; Brown, David (FAOKH) <David.Brown@fao.org>
Cc: 'SamAth Leng' <lengsamathy@gmail.com>
Subject: RE: Priority fields for vessel census database

Hi Dane,

Sorry to hear that matching the approved Excel data with the SQL data is so difficult. As suggested before, perhaps we should leave that for the moment. Please keep in mind the targets:

1. Fix data to match with official approved summary tables;
2. Agree on core fields to clean and eventually share the data outside of FIA in 2021; and,
3. Developing a web-interface to select, report and access individual registrations. That can also includes tools to effectively replicate the PDF summary tables, but should include all required tools to manage the database, lookup tables, user access etc…

Ad 1
I have taken a look through the DataDictionary file you sent yesterday. That is a bit confusing, but also useful to understand the main issues at hand. Also thanks for including a database diagram. Do we now agree to recommend to build a new SQL database based on the **1_AllV_15ct2019_Final_2020** excel data you sent before?

If yes:

- why do we need to link the Excel records with the existing SQL database for, if we are going to use the Excel data?
- if that is our official recommendation, this needs to be agreed with by Roitana (or is it Rattana?).

You told me before that **1_AllV_15ct2019_Final_2020** is Roitana's version, so what is the difference with **Official_All7552** (except that is using the questionnaire numbers as headings)?

I will not try to replicate your results, nor interfere with your process, I am sure it is fine. I just want to repeat that our target for this year is to allow reporting on at least 40% of the entries in the database. I understand that the main issue is to verify the content of individual records, but is there any way to assess what proportion of the records is clean and correct at this stage?

I fully understand that you need to keep Roitana happy, but that seems relatively straightforward if it is mainly to adjust where vessels are registered or which gears they operate, number of vessels with safety and communication equipment etc... My only concern, by fixing it to fit with the official summary results, do we incorrectly assign vessels to the wrong province or (length) category? That is where Roitana needs to be involved in making a decision.

Obviously, the real challenge is trying to verify the details for each record, but this may well be impossible based on the currently available data, due to data recording/entry and conversion errors, until the records are verified during licensing.

Ad 2
I think getting agreement with Roitana on the priority fields is a high priority, officially we need to give FIA, FIAC and WMPT access rights to 40% of the data within 2020. That is not going to happen, so then we should focus on getting the interface done in preparation of getting all records cleaned (100%) by the end of 2021.

My proposed priority fields, are exactly that, a proposal. Do we need to have a discussion on that before presenting to Roitana?

Ad 3
What is the progress with developing the web interface, or even a simple reporting database? What software tool are you going to use? I assume it needs to be PHP, or can this be handled within the standard SQL GUI?

Best regards,
Theo

From: So, Dane (FAOKH) <Dane.So@fao.org>
Sent: 10 December 2020 11:54
To: Brown, David (FAOKH) <David.Brown@fao.org>
Cc: 'SamAth Leng' <lengsamathy@gmail.com>; dangleloo@gmail.com
Subject: Re: Priority fields for vessel census database

Dear David,

I am going to work Mr. Sam Ath at least a week to finalize the dataset those key prioritized the field name, data cleaning consistency with official statistics. We have just discussed this morning how to process. To make it faster since he is backup all knowledge during the census, and we will work closely to confirm, approval with Mr. Riotana

Regards,

Dane

From: So, Dane (FAOKH) <Dane.So@fao.org>
Sent: Thursday, December 10, 2020 10:32 AM
To: dangleloo@gmail.com <dangleloo@gmail.com>
Cc: Brown, David (FAOKH) <David.Brown@fao.org>; 'SamAth Leng' <lengsamathy@gmail.com>
Subject: Re: Priority fields for vessel census database

Dear Theo,

Thanks for the instruction which very helpful, yes, I do wildchar mainly for report filter.

The suggestion ID (recorder_id, questionnaire_id, annex_type) is not work, there is completely modify with Ratta Data Source, Record ID and QID.

the main ID potential at village level (Interview province, district, commune and village) with Census No and annex_type. In each village level.

Yes, i have based on Mr. Rattana excel basis

There is no manual or description of data dictionary from the database, only a file I got last time and use to share with you (Database guideline.pdf). as i mention during my assessment on database, there is no proper define fieldname, some field should be number/ integer or decimal, date, but most of them were defined as text format.

I have sent (attached again some of my updates) you also my version DataDictionary.xlsx sheet 'DataEditingOfficialAgg' where i compare results and list potential need to be fixed or editing which

needs to confirm from Mr. Rattana.

Now, I am trying to only the key priority field you suggest and discussed with Mr. Sam Ath to finalize them and work with some report tables.

Regards,

Dane

From: dangleloo@gmail.com <dangleloo@gmail.com>
Sent: Thursday, December 10, 2020 8:28 AM
To: So, Dane (FAOKH) <Dane.So@fao.org>
Cc: Brown, David (FAOKH) <David.Brown@fao.org>; 'SamAth Leng' <lengsamathy@gmail.com>
Subject: RE: Priority fields for vessel census database

Dear Dane,

Thanks for the match-up and the questionnaire. I expected the fieldnames to be without a problem, because they were copied from an Excel version of the SQL database.

As a general approach, you need to both link fields between the Excel and SQL versions of the data and the questionnaire question number, as you are doing. However, it also is important to have clear descriptions for all fields. Apparently the rCode field is a code number, but it is unclear what it represents, i.e. what it is a code for. Are there lookup tables that were used for fields, besides the ones for vessel owner and captain, e.g. for gears and rCode? Is there any data dictionary that describes what the field contains and the type of data expected, so a full description of the meaning of the fieldname and if it is a text, number, date etc… field? Some of the descriptions you include are not clear, so are you working on a final list with clearer explanation? Perhaps there is a survey manual, or documentation for the database?

Yes, I am aware of duplicates for the primary keys in the Excel data. I assume you are trying to match Roitana Excel data with the SQL database, so can you give me some numbers on how many records match for the main ID fields (recorder_id, questionnaire_id, annex_type)?

I am assuming we are using Roitana's Excel data as the starting point for our target this year (being able to report 40% of the records). So how far have you progressed making a reporting interface for producing some summary tables, similar as to those included in the PDF? I understand that to decide what data is correct, you want to be able to match the Excel with the SQL dataset, and then ask Roitana to make a decision, but if you need to match the PDF summary tables, you should already have prepared a comparison table by province for what matches and what doesn't. Is that something you can share, or are you still working on that?

As for trying to match records, I am not going to second-guess your approach to data cleaning. Trying to match owener/master names by village for records that cannot be matched using the main id fields (as listed above), makes sense. If names were changed, you need to start looking at more fields to decided which record is equivalent with the one in the SQL database. I normally use simple code to compare names, where I try to match different length of strings for names between lists, using wildcard characters. That works quite well, just need to manual check the results.

However, I also understand that the SQL data is not complete, or at least that the number of records between the cleaned Roitana a Excel sheet and other versions of the data is not the same. The problem is that you are being asked, not just to clean the database, which is relatively straightforward, but also to make sure that the SQL database returns the same summary tables are already officially endorsed by the DG. In other words, you are being asked to correct the data to match the agreed results.

As we discussed, focus on cleaning the data first, for the priority fields and where relevant also include the “?” fields. If I know how many can be matched right now and how many are giving you a problem, we can discuss further.

Please also discuss directly with Sam Ath on the priority fields, so David can discuss this approach with Roitana ASAP.

Best regards,
Theo

From: So, Dane (FAOKH) <Dane.So@fao.org>
Sent: 09 December 2020 12:33
To: dangleloo@gmail.com
Cc: Brown, David (FAOKH) <David.Brown@fao.org>; 'SamAth Leng' <lengsamathy@gmail.com>
Subject: Re: Priority fields for vessel census database

Dear Theo,

Thanks for identifying. I have done matching fieldname available official data sources and server as well as with editing field to consistency with questionnaire which makes easy to reference. you can see here, i have brought back compare with yours.

attached your both matching and questionnaires. Some identidy, it is hard to get from server information detail such IDNo etc due to mismatch key primary fieldname both dataset, the best approach short at village level and compare as manual name by name, ID, i have try to develop a script to cross several time but still not yet get a result.

I will prioritize those 'y'.

Please advise for an alternative solution or any additional

Regards,
Dane

From: dangleloo@gmail.com <dangleloo@gmail.com>
Sent: Wednesday, December 9, 2020 10:19 AM
To: So, Dane (FAOKH) <Dane.So@fao.org>
Cc: Brown, David (FAOKH) <David.Brown@fao.org>; 'SamAth Leng' <lengsamathy@gmail.com>
Subject: Priority fields for vessel census database

Dear Dane,

Please find attached my list of priority fields. I am not a marine fisheries expert, but I have tried to find a balance between what would be needed for the core vessel census data and what seems more relevant for the data collected for the license, logbook and catch monitoring databases. As we discussed, for the core vessel registration we need much less, but I understand this was trying to collect ‘everything’ relevant.

I have indicated field I think are required with “Y”, for fields I am not certain I have added a “?”.

Perhaps Sam Ath, can also take a look at the list and advise what information makes sense to include in the core data set?

I also noticed that some information is missing, for example: construction year, country of build and radio call sign were not recorded. Other fields may be recorded, but I don’t see them in the main table, e.g. mobile phone number, business registration number or tax number (for any companies running multiple vessels).

After we agree, we can discuss this with Roitana.

Let me know if there is anything else.

Best regards,
Theo