OVERVIEW AND CURRENT STATUS OF THE RTTP-IO DATA INCLUDING INDICATIONS OF ITS STATE OF READINESS AND LIMITATIONS

Jean-Pierre Hallier[1]

This document is made to give an idea of the data collected by the RTTP-IO and it present status in term of quality control and validation.

Data collected are either related to the tagging operations or to the recovery. All tagging data were collected at sea on board the two vessels chartered by the project while recovery data were collected either at sea or on land in many different contexts which lead to a great heterogeneity of their quality.
The RTTP-IO also suffered a lot of shortage of manpower especially up to February 2007. This lack of staff has mostly affected the recovery aspects of the project. Hopefully recoveries were slow to come at the beginning of the project and we received great support from the IOTC.

1. TAGGING DATA

All data collected at sea were entered onto laptop at the end of each day in an ACCESS database by the Cruise Leaders (CLs), Tagging Technician (CTTs) and Regional Tagging Technician (RTTs). For the tagging data, it was always the tagger who entered his own tagging data. While tagging, data were first registered on a digital recorder that was hanging down the neck of the tagger. Then the voice recordings were transcribed on paper sheets before being entered into the ACCESS Database. Apart from tagging the other data collected were collected and registered in the following tables:

- The cruise table (periods at sea were divided into cruises);
- The vessels' activities;
- The bait hauls (baiting activities);
- The bait utilization (amounts of bait available on board);
- The schools sighted as well as other sightings like logs, FADs, whales or whales sharks;
- The catch-on-deck (any commercial fishing was forbidden, therefore all fish coming on board were registered);
- The length frequencies of some on the fish on-deck;
- The biological data for some of the three tuna species on deck;

While entering data into the database, ACCESS own system was tracking inconsistencies such as duplicate tag numbers or data which were not those pre-registered.
At the end of each day or the morning of the next following day, all data collected the previous day were E-mailed as CSV files to the Chief Coordinator and to the IOTC. These data have already gone through rough checks of possible mistakes by the Cruise Leaders on board.
Once received in Seychelles, the CC made a first rapid check of the data and reverted back to the vessels for immediate corrections on the database if necessary.

Once tagging was over, different ACCESS queries search for possible mistakes or inconsistencies into the TagRelease table and corrections were made whenever necessary. Checking of the other tables is not finalized yet and will be done in the next future.

[1] Chief Coordinator of the RTTP-IO, c/o IOTC, P.O. Box 1011, Victoria, Seychelles E-mail : jph@iotc.org

Some recoveries highlight some possible mistakes into the tag release data. Tagging is a fast operation and we have tagged very often mixed schools of the three species. Furthermore, in Tanzania, small YFT and BET were sometimes very difficult to distinguish even as fresh fish. While recording on digital recorders, many incidents can occur that can affect the data registered or their quality. Other errors can take place while voice recorded data are transcribed on paper and then while they are entered into the database. It does not mean that the database is full of errors; it just means that we are aware of these possible mistakes and we are prepared to correct them when it is possible.

When some fish are recovered it turns out that the recovery data does not fit with the tagging data:

- The species is different;
- The lengths between tagging and recovery do not match (length at recovery is smaller than length at tagging or growth is too high or too low);
- Date of recovery is anterior to tagging date;
- Etc...

The case of differences between tagging and recovery data is detailed in the Recovery data paragraph of this section on the quality control. It should be noted that all species changes are documented in the tag database.
There are 16-17 different data recorded for every fish tagged plus 3 other essential data coming from the sighting table: Date and position of tagging and the school association. There are 7 other secondary data into the sighting table.

In ACCESS we have 22 queries which are searching for possible mistakes or data problems related to the essential data. They have been run and corrections made. Therefore very little mistakes remained in the tagging data apart from those ones which can be revealed through the recoveries.

2. RECOVERY DATA

As mentioned previously the recovery data suffered at the beginning (until February 2007) because of staff shortage. This point needs to be considered when one wants to appreciate the quality of the data.

2.1. Recovery data collection

All recovery data are first written on the Tag Recovery Form (TRF – figure 1). On the TRF, there are 21 essential data to collect and 11 secondary ones. The fact, that most recoveries occur in port while purse seiners are unloading or transhipping their catch, increases the number of data to collect. In this latter situation we don't have the date of catch but we have to record the vessel name, the date of unloading and the well number(s) from where the recovered fish was found. Later with these data we can match them with the log book data of the purse seiners to obtain possible date(s) and position(s).

The way recovery data were collected changed with time and with location. In Seychelles, at the beginning when we have only one person in charge of the recovery collection (a Tag Recovery Officer - TRO), either the stevedores bring the fish to the IOTC office that was conveniently situated inside the fishing port or they bring a string for the length of the fish together with the tag and the other data. This system proved very quickly to be inadequate for all essential data such as species, length, vessel name and vessel well. Therefore with the agreement of the PS captains we asked the stevedores to systematically bring the fish to the IOTC office where at least we can ascertain the species and the length. At that time almost all recovery fish were small (either SKJ or juvenile YFT and BET); but later on it became

unpractical as some fish get bigger. But bringing the fish to the office, independently of the messy situation it creates, did not solve the data collection for vessels' names and well numbers. When the number of recoveries registered per month bypassed 500 or 600 hundred, not only the RTTP TRO was full time busy but also the IOTC TRA, Julien Million, as well as the RTTP PTRO, with some occasional help from other IOTC staff as well as RTTP CC and AFO. But when we reached more than 1000 monthly recoveries in December 2006, the available staff could not cope. Hopefully we manage to secure another TRO in February 2007 and 4 Tag Recovery Assistants (TRA) in March 2007. With the arrival of this team we change our data collection procedure. The tag seeding experiment have shown that vessels'names sometimes but more often vessels'well numbers were not properly reported by the stevedores. Furthermore, with the large number of fish as well as the increased proportion of larger fish some PS captains became reluctant in letting the stevedores bringing the fish to IOTC office. Not only the fish was lost for the vessel but the vessel often did not see the stevedore for some times loosing the paid manpower. Therefore it was decided to send our Tag Recovery Team as close as possible to the finding of the tagged tuna. We divided the team in three groups of 2 staff each. The stevedores were instructed to keep the fish next to the well where it was discovered and two groups toured the PS in port or at anchor to collect all the appropriate data; the last group being in the office for paying rewards, doing the necessary photocopies as justification of payment and collecting the PS names in port and wells which were unloaded every half day. This system proved efficient and is still into place today.

We have also introduced the collection of weight together with the length at the beginning of 2006 on a relatively small level until mid-2006 when we generalize the system. Later on the relationship FL/WT will help to tackle some measurement errors and it gives more strength to the length and weight data when they match well.

2.2. Recovery quality control

The TRF is filled on the spot. Then they are gathered in the office where one of the TRO screens them quickly to spot any unfilled field and any obvious mistakes. Once this is done, the procedure differs a little bit if data are collected in Seychelles by our Tag Recovery Team (TRT) or in Seychelles at the cannery, at sea or outside Seychelles by our Tag Recovery platforms disseminated in many different places.
If collected at sea, in Seychelles cannery or outside of Seychelles they are first checked by the PTRO, Teresa Athayde, before they are passed on to the data entry secretary, Betty Honore.
If collected in Seychelles by our TRT a subsample of the TRF are revised daily by the PTRO before they are all given to Betty for data entry. Quarterly the PTRO conducts some general check on the quality of the data. She has defined several parameters and she checked their occurrence (% among recoveries). These are length and weight reliability codes, species, date and position. As for the Tagging database, ACCESS checked for appropriate data when data are entered. It is then almost impossible to duplicate tag numbers. However double tags, especially when the first tag was lost, causes us some problem. Once entered the TRF is stamped by Betty to ascertain its entry inside the database together with the date. Any TRF checked by the PTRO or by the CC are equally stamped to register the checking operation.
TROs and TRAs also participate to this quality control, for instance they compare rapidly tagging data (species, tagging date, length and type of tag) to the recovery data. If they notice some discrepancies (species difference, lengths not adequate, etc...) they pass on the TRF to the PTRO or the CC for a closer look to the recovery data and the corresponding tagging data and decide eventually on some corrections. Generally species discrepancies will be left unattended in order to deal later with them (see below). Whenever necessary Betty will enter the correction and then scan the forms or pass them on to someone else for scanning (all TRF are scanned). At the end the TRF is archived in files and stored in filling cabinet (we already have 8 filing cabinets full of tags).

2.3. Data validation and storage

A total of 29 queries in ACCESS were designed by the CC and written by the IT Administrator of the project. They are regularly run by the CC on the recovery database for the detection of potential errors. Some of these errors can be easily attributed either to tagging or to recovery and data are immediately corrected into the corresponding databases.
I will not list here the entire different searches made by these queries but we have tried to figure out all the possible mistakes, all missing data and all inconsistencies in the data. Query results are not always resulting in corrections either because it is not necessary or because we don't have the information necessary for the correction. Furthermore, some problems are more difficult and require more work; this is the case for the species discrepancy.

2.4. The species misidentification problem

A special query was designed to tackle this problem in order to help the correction process. The query gives some information on the length/weight relationship, the growth (in cm/month), the tagger, the recorder of the recovery data, the dates, etc... It was not possible to design an automatic system for species discrepancy correction because too many factors need to be considered some of them more qualitative than quantitative. Therefore the 1360 species discrepancies recoded so far were treated individually.
The different factors taken into consideration are as follows:

- The species is U (unknown) at tagging or at recovery; the known species is applied to the U species; if both species at tagging and recovery are U it remains U in the final database;
- The growth is quite different between species;
- The length/weight relationship also different between species at least for certain size ranges;
- Who has seen the fish: one of our Tag Recovery Staff or another person (stevedores, cannery worker or fisherman). We know that fishermen and often stevedores as well as cannery workers don't recognize the existence of BET for small fish (sometimes up to fish of 10 kg or 70-80 cm). The grouping of small tuna whatever the species under the name SKJ is also not rare.
- The tagger is considered as some are more prone for misidentification; for instance a new tagger not yet experienced;
- The tagging period as we know some periods are more prone to misidentification; for instance when mixed school show up after a long period dominated by SKJ or YFT;
- The person who collect the recovery data;
- The recovery period; for instance we were short of staff until February 2007 and therefore errors are more likely to occur during this period;
- The species tagged in the corresponding school; for instance if a YFT is given at recovery for a school where only SKJ were tagged or during a period when this species was not tagged the species at recovery is wrong;
- The species reliability at tagging; for instance if it is listed as unsure then we will give the preference to the recovery species if other parameters also support the recovery species;
- If any problem with length we also check if the fish is noted as damaged, bent, cocked or any particularity which can explain a bad length. Normally the length reliability of this fish should have been noted "Bad";
- We are also checking the paper trace at tagging to detect a possible typing error at entering the data into the database (we can, but we have not done it yet, go back to the voice recording to check if an error occurred at this stage);

- We are also checking the Tag Recovery Form (TRF) on which data were written to check for a possible data entry error. This can be done either by looking at the scan of the TRF (they are all scanned) or directly to the TRF (we have heaps of filing cabinets with arch files full of them).

As mentioned all TRF are scanned. When corrections are made they are reported on the TRF that needs to be re-scanned. For species changes, they are all coded and entered into the Recovery database even if the species change is made in the tagging database. Of course when the species change is affecting the tagging this change code is also reported into the tagging database. The coding system put into place gives the possibility when data are analyzed either to discard these recoveries or to use them according to the corrections or even to go back to the original species. The coding is made of:

- CH for change of species;
- R or T to tell if the change concerns the recovery species (r) or the tagging species (t);
- Two letters regarding the species, the first is the original species and the second the final one (species are coded Y, B & S).

Table 1 summarizes the species changes. It gives for each species the frequency of each different change and some synthetic numbers assessing the importance of those changes. It should be noted that these 1360 species misidentification represent 5.6% of all recoveries (5.7% if you add the 41 unsolved). However 1.3% are related to unknown species (U) at tagging or at recovery which are changed to the known species (256 at recovery and 52 at tagging). Therefore strictly speaking the species misidentifications account for 4.3% of all recovered fish: 1.9% at recovery and 2.4% at tagging. Considering that (1) nearly 80% of all fish were tagged on mixed schools of yellowfin, bigeye and skipjack and (2) 50% of the recoveries are YFT and BET, this proportion of misidentification can be considered as low and acceptable.

There are still a dozen recoveries with unknown species mostly because the species was unknown at tagging and at recovery.
When parameters considered are not sufficient to support a change of species at tagging or a t recovery, data are left unchanged.

2.5. Different parameters illustrating the recovery data validation

After the species, the other important parameters are the length and weight and their reliability, the fishing vessel and the well numbers for purse seiners, the associated dates and positions.
Lengths are taken as much as possible in FL; we avoid taking length in FDL for not adding more uncertainties to the data as all analysis are done using FL. Weights are taken as much as possible. Length reliability is good, bad or unknown. 96.3% of all recoveries have length and 73% have weights. Figure 2 is giving the distribution of the percentage of length reliability with time; figure 3 is for weights reliability. Figure 4 combined the two previous information by giving the percentage of recoveries with good lengths and weights.

Globally since mid-2006, the situation is good especially when you keep in mind the fact that only 13% of the recoveries took place before mid-2006. However the qualification good as length reliability does not mean that no error exists on the length, only what the recorder of the information thinks about his measurement or that he trusts the person who provided the length.

We still have 967 recoveries with negative length increments (Length at recovery – length at tagging). We still need to have a look at these errors but it is time consuming and they represent only 4% of all recoveries. Therefore it was not among our priority for the WPTDA.

2.6. Attribution of possible dates and positions to recovery made on PS in port or in canneries

Among the recoveries, 24% are related to a known date of catch and position (recoveries detected at sea), the remaining are associated to one (or several) vessel(s) and to one (or several) well number(s). Using the PS logbooks, we can link this information to one (or several) sets with their corresponding date(s) and position(s); ending with possible date(s) and position(s). This task was realized on most of the concerned recoveries. This is done by running special software called "Data Editor" that brings together recovery data and PS logbooks. This is not an automatic system it requires a manual decision for each recovery. With the help of the Data editor, possible date(s) and position(s) of sets have been associated to recoveries from PS unloading in port or from reefers or canneries.

Of the 23,472 recoveries from the purse seine fleet, 5,512 have a known date of recovery related to a discovery of the fish at sea. The 17,560 remaining recoveries were found in wells at unloading or transhipping operations or in the canneries (cold store, processing plan). 16,412 of them have been "edited" which means that the vessel and well numbers registered were matched to the log books of purse seiners in order to assign to this recovery the dates and the positions related to the fish present in the wells. Therefore 1,548 have not been edited yet (generally recent recoveries). A status is assigned to the outcome of this process which describes how successful it was:

- Complete = there was no problem in matching the data;
- Incomplete = some discrepancies exist between the vessel-well information from the recovery and the logbook;
- Missing data = Data were missing in the log book to complete the process;
- No reliable convergence = we could not match the two sets of data together;
- Unknown origin = for fish with so little data at recovery than we cannot related to any log book.

The distribution (in percentage) of the status assigned to the recoveries is given in figure 5. Nearly 80% of the data can be associated to dates and positions of sets from purse seiners; 6.7% are associated to purse seine data which do not match completely with the recovery data and only 16.3% cannot be associated to PS sets.

The recoveries with the edit status complete and incomplete (77% of all recoveries) are associated to different numbers of sets. The distribution of the frequency of these numbers of sets is given in figure 6.
Large numbers of sets correspond to situations where several different wells are associated to the recovery or even several vessels (sometimes the case with recoveries done on board reefers or in canneries). It should be noted that more than 50% of these recoveries are associated to 6 different sets or to less than 6 sets; a reasonable number of sets (reasonable dispersion in time).
We assess the dispersion in time of the different sets using the duration between the lowest set day and the highest set day. The resulting frequencies are given in figure 7. Altogether 75% of the recoveries are associated to sets that are 7 days apart or less than 7 days apart.

We have not done yet an assessment of the corresponding dispersion of the sets of each recovery in space.

These dispersion parameters will in some ways give an idea of the degree of confidence one can put into an average date and average position for these recoveries. To better assess the quality of these dates and positions we still have to give some weights to the different sets of a given recovery.

2.7. Different dates associated to the recoveries

All recoveries possess as Date at least the Date of Return but they also can have a Date Found (often the same as the Date of Return), a Date of Catch (when found at sea; in this case Date of Catch and Date Found are the same) a Date of Unloading. Finally we can have a Date of Unloading for the reefer when the tag is found on a reefer. When associated to a vessel and wells we will also have the possible dates from the data editor. By degree of precision the dates are ranked as follows: Date of Catch, Possible dates of sets, Date Found, Date of Unloading, Date of Return.

3. VALIDATION AND CALIBRATION YET TO BE DONE

The validation process is an on-going operation which will last until the end of the RTTP-IO and then carried on by the IOTC. The same principle already in place will be applied through out the new data. The main points still to be carried on are:

- The resolution of some recovery position on land (most have already been corrected but some remain);
- The refinement of the PS data editor to avoid negative time-at-liberty. The PS trips at sea last at the most 45 days, when a recovery is made not too far from the tagging date and it is discovered in port the well(s) link to this recovery might contain sets with fishing dates anterior to the tagging date. It should not be possible to attribute these sets to the recovery. For the moment, this aspect is not taken into consideration; this will be corrected;
- We might attributed some parameters to assess the dispersion in time and space of the different sets linked to a recovery ; this objective is to give an idea of the degree of confidence which is attached to recovery average date and average position;
- Following the same principle, it will also be necessary to give some weights to the different sets link to a given recovery;
- Negative length increments have not yet been systematically checked to eventually correct them; we will have to do it.

These are the aspects which we will work on before the next October WPTT.

4. CONCLUSIONS

The data quality control as well as the validation and calibration of the data are very time consuming because it need to be done with great caution. There is no point in speeding up this process with the risk of replacing wrong data with false ones. Most of this work can only be carried out by staff who know very well the data (tagging and recovery) and within the RTTP-IO, staff with this quality is limited to two persons: the PTRO and the CC. For some of the aspects we also get the help of the Seychelles Tag Recovery Team. Therefore the scientific community must be patient. We tried our best for this meeting but the time was missing to end up with the cleanest databases as possible. During the intense tagging and recoveries of 2007 we were forced to put aside a lot of tasks including most of the data validation and this is only recently that we see the end of the backlog accumulated for many months. We are still busy with recoveries and we want to put more emphasis on recoveries from the longline fleet. But we think the data we are presenting now show well that we are moving in the right direction on this data quality aspect. The work of the all team is to be praised for the work already done

Table 1: Distribution of the species misidentifications and the way they were dealt with (not represented here are 41 misidentifications that cannot be solved).

	YFT	BET	SKJ	Total
CHrBS			38	38
CHrBY	102			102
CHrSB		64		64
CHrSY	73			73
CHrUB		23		23
CHrUS			121	121
CHrUY	112			112
CHrYB	1	218		219
CHrYS			98	98
CHtBS			17	17
CHtBY	150	1		151
CHtSB		13		13
CHtSY	57			57
CHtUB		12		12
CHtUS			20	20
CHtUY	20			20
CHtYB		175		175
CHtYS			45	45
OVERALL				
Change of Species	515	506	339	1360
% change	6.3	11.0	2.9	5.6
AT TAGGING				
Change at tagging	227	201	82	510
change at tagging w/o U	207	189	62	458
% change overall	2.8	4.4	0.7	2.1
% change w/o U	2.6	4.1	0.5	1.9
AT RECOVERY				
Change at recovery	288	305	257	850
Change at recovery w/o U	176	282	136	594
% change overall	6.3	11.0	2.9	5.6
% change w/o U	2.2	6.1	1.2	2.4
Total recoveries	*8114*	*4605*	*11735*	*24454*

Figure 1: Tag Recovery Form used by the RTTP-IO to collect data related to recovery

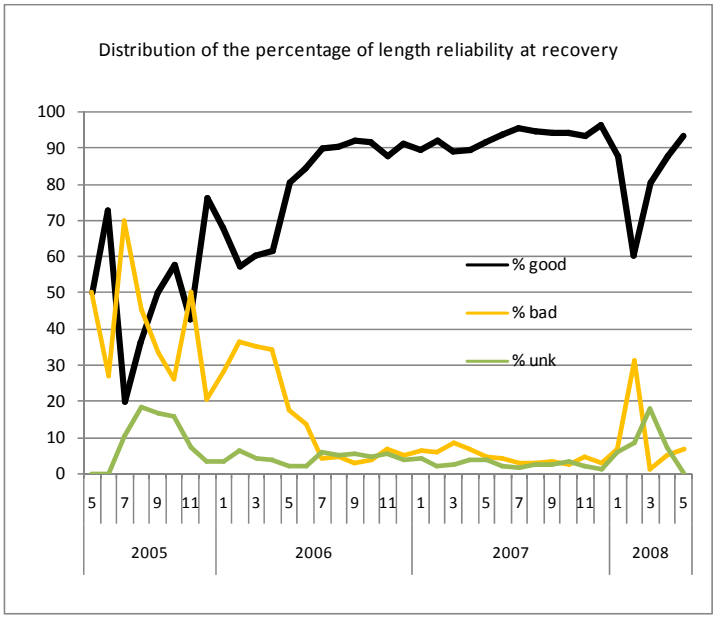

Figure 2: Status of the length reliability in percentage of the total number of recoveries with length

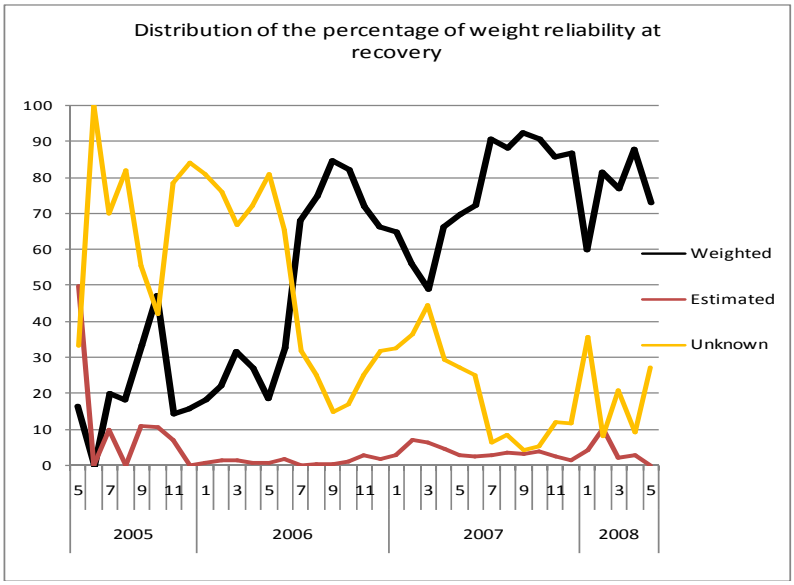

Figure 3: Status of the weight reliability in percentage of the total number of recoveries with weight

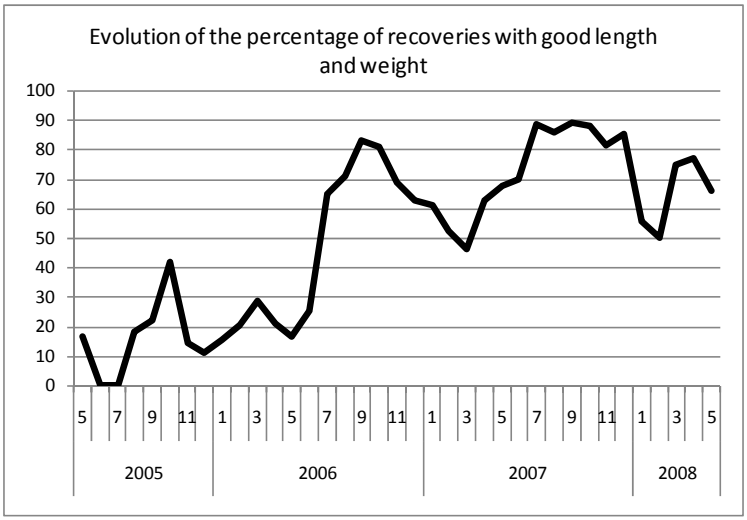

Figure 4: Percentage of recoveries with good lengths and weights

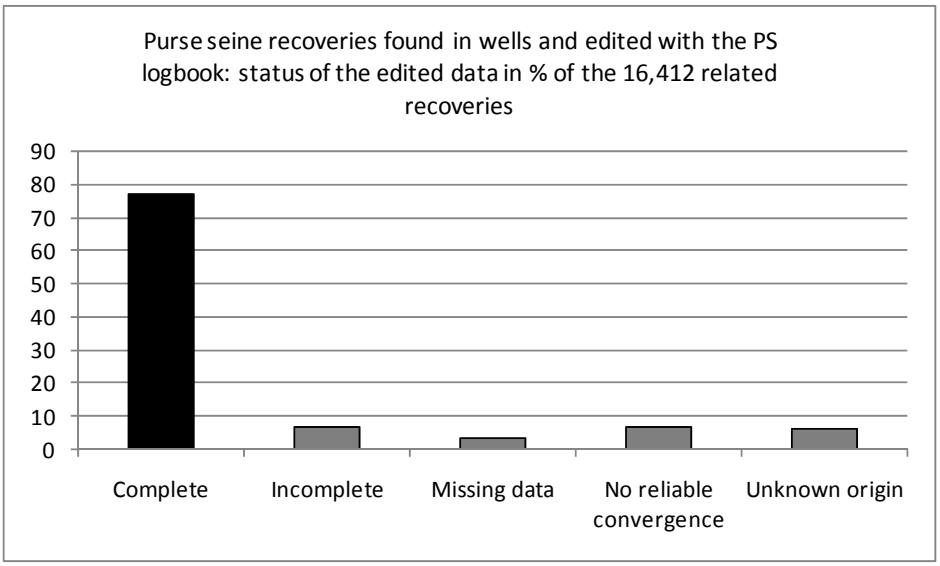

Figure 5: Status of the recoveries edited for possible date(s) and position(s) of recovery

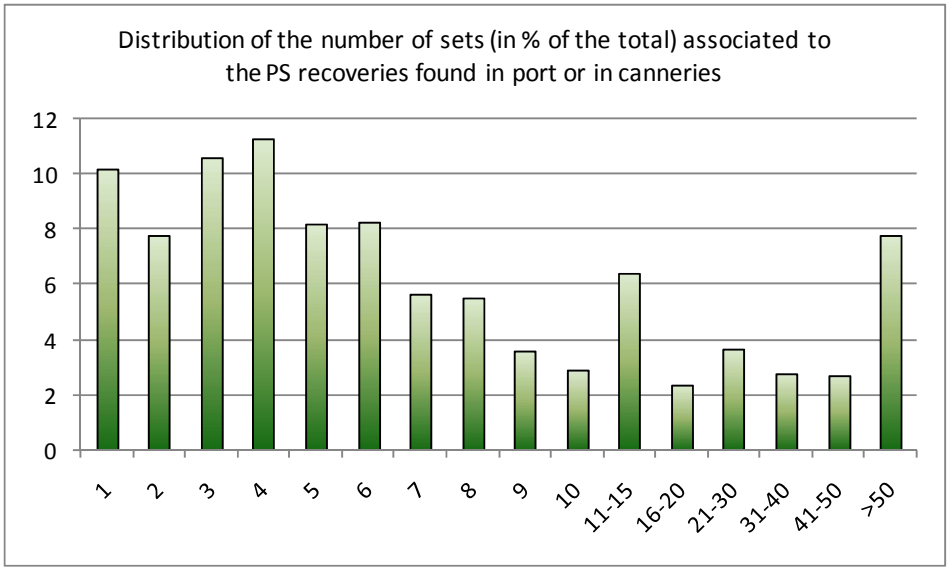

Figure 6: Distribution of the recoveries according to the number of sets linked to them when they are edited with PS logbooks

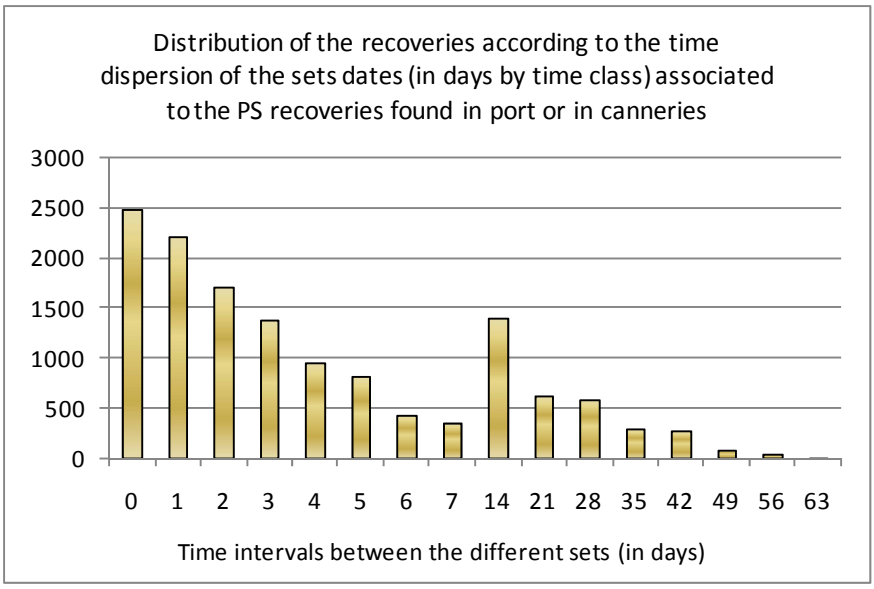

Figure 7: Distribution of the recoveries according to the dispersion in time of the different sets linked to the recovery by the data editor using PS logbooks