DEEPFISHMAN
Management And Monitoring Of Deep-sea Fisheries And Stocks

Project number: 227390

Small or medium scale focused research action
Topic: FP7-KBBE-2008-1-4-02 (Deepsea fisheries management)

DELIVERABLE D1.3

Title: Data format and protocole

Due date of deliverable: month 12

Actual submission date: 06 July (month 16)

Start date of the project: April 1st, 2009   **Duration :** 36 months

Organization Name of lead coordinator: Ifremer

Dissemination Level: PP

Date: 06 July 2010

Research project 2009-2012 supported by the European Union,
Seventh Frame Work Programme

1. Introduction

This document describes how data will be handled in Deepfishman. Beyond the technical aspects the following rules should be drive the activity of partners in the projects:

2. Scale

Deepfishman works at a Case Study scale and data should be collected and compiled at Case study scale. Nevertheless, the boundaries of case studies may be unclear in some cases. For a simple example, biological parameters of a given species may not have been estimated in the geographical area of a given case study where the species is however fished. In this case partners should identify which biological parameters have been used where they come from and provide them with all useful comments as metadata.

3. Metadata

Metadata are all information useful to use data. Metadata should include all descriptions of variables included in tables/databases. This is primarily full name of variables and units. For geographical position it should be specified whether data are decimal degrees; degree, minutes and hundredths or degree, minute and seconds.

Metadata should also include the origin of the data, how of was compiled, by whom and any information on reliability and things to know to use data properly, date of release, information about how to acknowledge/cite the data set.

When standard formats are used, metadata are simplify and may be a simple statement such as:
- o (for sampling, landings and effort data from commercial fisheries) the format in this data file is in agreement with DCF, format, units and variable names from ICES (2009) were used (http://deepfishman.hafro.is/doku.php?id=partners:work_package_1#data_format_and_protocole);
- o (for survey data) this dataset is provided in DATRAS format (see: http://datras.ices.dk/Home/Descriptions.aspx)
- o (landing effort) This dataset is in ICES InterCatch format (http://deepfishman.hafro.is/doku.php?id=partners:work_package_1#data_format_and_protocole)

If minor deviation from standard format occur, these should be described as metadata.

If a dataset or a database is provided under a format defined, in a previous, EU, International or national project, the Metadata file can be the format description in this previous project.

Metadata should include any comment on completeness and reliability of data.
For example, a file of landings data can be provided with comments such as "*data are reliable from year 1999 to 2008, year 2009 is provisional, year 1998 is known to be incomplete, reliability of data before 1998 could not be checked but it is believe that some landings are missing due to [....], These missing data may represented x % of reported landings.*"
Such information is essential to the project as uncertainties about data and missing data may be accounted for in some modeling.
Small metadata files should be provided both in Case study report and as files provided together with the data, if a report describing the format is available, the Case study report should only include the reference and the report should be provided together with the data.
Data exchange formats should be uploaded to the wiki (section partners/ work package 1/ Data format and protocols).
 When metadata files are only a few lines of text for every data file. It is suggested to use understandable file names such as: French_DW_fleet.csv and French_DW_fleet_metadata.txt (or .doc) for the data file and metadata text respectively.
In the case of access database, part of the metadata can be include as variable description. Nevertheless, it is recommended to also copy variable descriptions it a text file.
For examples of data and metadata files of the French fleet in CS2 see CS2_fleet_data.zip

4. Completeness

Many available datasets are likely to be incomplete. Incompleteness is not a reason not to provide data at all. Partners should provide all available data. Reasons for incompleteness, if known, should be specified in the Metadata.
Data that have not been collected for a given Case Study, fleet or species should be identified in the Case Study report.
For numeric data, it should be specified in the metadata whether they are complete or not. For incomplete data, the difference between "not available" and "true zero" should be made clear. For example, if a list of vessel is provided as fleet data and a file of landing by vessels as landings data there should be a zero where the vessel did not land the species and a "NA" (Not Available) if landings data for that vessel were not provided.
If more complete data are expected to become available in the near future, readily available (incomplete) data should however be provided. Over time more complete data will be made available to the project as a new release.

5. Alternative data format

For commercial fisheries data, there are two main data formats. Data for samplings, landings and effort data defined under the EU COST project and ICES InterCatch format. The description of these two standard formats are available on the project wiki (http://deepfishman.hafro.is/), zone partners, WP1.

Any other data format that have been defined under DCF[1] (formerly DCR) or EU COST project, other EU project are acceptable to Deepfishman, provided data format are described in the metadata. Difference from current exchange format available on the DEEPFISHMAN WIKI should be described with caution.

Similarly, survey data may be provided under DATRAS format or MEDITS format with a one sentence metadata file specifying the format of the data (e.g.: this data from the survey XXXXX is provided under the MEDITS format (see format on the Deepfishman wiki and http://www.sibm.it/SITO%20MEDITS/principalereports.htm for a fuller account). In case of the use of any other format, this format should be fully documented.

It is recognized that all partners in the project may not have a full expertise in data management. Therefore, if data are available under different formats that those required here and manpower to do format conversion is not available at partner level format, data should be made available under their current format. In this case, metadata are particularly essential.

6. Aggregation/anonymisation

There are EU and national rules about the protection of electronic data on individual persons and companies. Only aggregated and anonymised data will be distributed beyond of the

---

[1] DCF: *Data Format Framework, COMMISSION DECISION of 6 November 2008 adopting a multiannual Community programme pursuant to Council Regulation (EC) No 199/2008 establishing a Community framework for the collection, management and use of data in the fisheries sector and support for scientific advice regarding the common fisheries policy*

project and used of disaggregated data within the project will remain under the control and responsibility of Case study leaders. Nevertheless, case studies are required to collect as much as possible disaggregated data and store then at partner level. Different aggregation method may then be required according to data processing to carry out.

For example, effort, landings and data identifying individual vessels cannot be distributed out of a given institute (bidden by EU and/or national confidentiality rules) the formal agreement of vessel owners. This applies to database EFFALO, TACSAT, ECOENQ1: and ECOENQ2 listed below.

Partner institutes are required to assemble their data under these formats. These disaggregated data will be used only under their control (during workshops or using processing routines circulated by WP leaders).

7. Source of data in Deepfishman

Most of the data in the project will be collected, assembled and made available at Case study level. Case studies should provide one or several database, data archive, datafile to the project. During the duration of the project data will be available on the Deepfishman website: (http://www.ifremer.fr/deepfishman) and/or the Deepfishman WIKI, in Case study pages: (http://deepfishman.hafro.is/doku.php?id=partners:partners). Data aggregation/anonymisation whenever necessary will be done whenever necessary. After the project, part of the data that are not subject to confidentiality will be stored on a web-based library (e.g. PANGEA http://www.pangea.de). However, the project will also make use of data where individuals/companies are identified and such data can be made available without appropriate aggregation at e.g. national and/or marine region level.

8. Data base –data archive

Data files can be provided as access database, zip archive or text file. In all case metadata should be provided together with data themselves.

8.1. Access databases.

Access database are comprised of several data tables. In a structured access database all tables are connected to each other by relational links. Nevertheless access data bases may also be use to store data without organizing link between data tables.

For example, a simple database could be a table of fleet data and a catch and effort data. The fleet data would include the following variables

Community fleet register number
Vessel number
Port
Overall length
Power

The catch and effort data would be total catch and total effort in days at sea of vessels in the fleet table, but vessels are no identified because catch and effort of several vessels in the same day and statistical rectangle are cumulated.

Year

Month
Day
Rectangle
Total catch
Catch species 1
Catch species 1
Catch species 1
Catch species 1
....
Number of days at sea
Number of days fishing

In this case there is no common variable between the two tables, no logical link can be establish between them they can still be included in the same access database with no relationship. Please note this is a fictitious example.

8.2. ZIP archive

The same data could be included in separated files (excel, text) and stored in a single zip archive.

For text files, it is recommended to use CSV (comma separated variables) format that is easily read by programming languages and statistical analysis software such as R.
It is particularly important in zip files that data file and metadata are included. Please see the rule for filenames proposed in section *3 Metadata.*

8.3. Charts

Data for charts can be stored as shapefiles. Data under this format data can be by GIS software such as ARCGIS and the R package PBSmapping. For charts also all relevant metadata should be included.

8.4. Datasets

Depending on the amount of data, all data from a given Case Study may be included in a single database of several may be produced (e.g. fleet, landings and effort data; survey data; biological parameters; VMS).

9. Accessibility

Database or archives will be made available to project partners on the partner area of the website (http://www.ifremer.fr/deepfishman) or deepfishman wiki. Where applicable, only aggregated/anonymised data will be distributed.

10. Update

During the course of the project data will need updating.

Updating may become necessary when (i) new data become available, (i) errors have been detected by Case Study orWorkpackage scientist as data processing often reveal inconsistencies that require coming back to the data and correcting.

Updating should be validated by the Case Study leader. If project partners find problems with data from one Case Study, they should get in touch with the case study leader, analyse the problem and if possible correct the data and/or specify the problem in the metadata. if it is possible to identifiy which reccord in data file are dubious a pointer can be added to identify with records are reliable or not or the classify data reliability (e.g. a scale from 1-poor to 5–good). Information on the coding of the pointer should be added in the metadata.

11. Required variables per type of data

11.1. Fishery data

11.1.1. Fleet data

Fleet data should include as many of the characteristic of studied fleets. In particular data from the following table should be provided. For EU fleet, the Common Fleet Register (CFR: see http://ec.europa.eu/fisheries/fleet/index.cfm) format should be primarily used.

If not available, data consistent with the fleet description required in the Case Study report should be compiled.

Table 11.1.1a. Fleet data to be provide (From the template for Case study report)

Nationality
Gear type
Fleet ID for use in other fleet table
Fishery type:-target/mixed fishery/bycatch
If mixed or bycatch what are other or target spp?
Number of vessels
Large scale or artisanal
Time period

Table 11.1.1b Data for vessels in the fleet

Fleet ID (from table above)
Country of registration (flag country of the vessel)
Common Fleet Register number (code of the vessel in CFR, see commission regulation (EC) No 26/2004 of 30 December 2003 on the Community fishing fleet register
National code of the vessel
Other vessel identifier (if any)
Data of the vessels as listed in commission regulation (EC) No 26/2004 (CFR regulation) 36 variable are identified in the regulation. The provision of these is mandatory under DCR. Non EU partners and Case study will provide similar information as much as possible
Other vessels data relevant to Deepfishman

11.1.2. Catch and effort data

Catch and effort data can be provided under COST format for EU countries (ICES 2009), ICES InterCatch format or any other format properly described in a metadata file.
Other data such as input format to standard assessment routines (e.g. FLR, gadget) can also be used for a given stock, fleet or area provided it is fully described in a metadata file.

11.1.3. VMS data

Raw VMS data, include the following information: Vessel CFR code, date, time, Latitude, longitude, vessel heading.

VMS data might be provided to DEEPFISHMAN with appropriate aggregation and anonymisation. For anonymisation, vessels data (power, length gear, vessel type) should first be associated to the VMS data, them vessels codes may be replaced by a running number. Data can then be aggregated by geographical units. It is highly recommended that a higher geographical resolution that ICES rectangles is used. Rectangles of 5 to 10 nautical miles seem an appropriate spatial aggregation level. Any other may be used, justification should be provided. Filtering between fishing and steaming VMS records should be applied and data of aggregated total VMS records (all VMS record for the Deep-water fishing fleet or area) and aggregated VMS fishing effort (aggregating VMS records of VMS data estimated to represent fishing activity based upon the filters used) should be provided.

Alternatively any other estimates of deep-water fishing effort based upon VMS data should be made available (data file and corresponding metadata) to DEEPFISHMAN.

11.1.4. Observer data

On-board observation data should primarily by provided according to COST format (see: ICES, 2009. Definition of Standard Data-Exchange Format for Sampling, Landings, and Effort Data from Commercial Fisheries ICES Cooperative Research Report, N° 296, July 2009. 48pp.

Nevertheless, data set that were collected and stored under different format (i.e. archiv e data or data collected in the EU out of the Data Collection Framework) may not be easily converted to COST format. In this case the original format may be kept, it is then essential to provide appropriate metadata, together with the data. See below annex 1 et 2, examples of data and metadata.

For observer data, section 6. Aggregation/anonymisation above might need to be considered and appropriate aggregation and anonymisation might be done if required.

11.1.5. Abundance indices

In addition of catch and effort, on-board observation, survey and VMS data any time series of abundance indices should be made avaible together with the estimation method and details on data used as metadata.

11.1.6. Data form fishers, fishing organisation and other stakeholders

These data may be in Ad Hoc format and will require specific anonymistion/aggregation. There is no standard format for such data. Case study should provided avaible data with appropriate anonymisation, aggregation and metadata. Restriction for dissemination beyond the project may apply (if so specify in the metadata).

Some data may also be available to a given institute but not to the project. In this case, project partners should report in Case Study report which data has been provided by stakeholders, describe the data, describe what use of the data has been made or will be made and specify confidentialy reservation to this data. Such data will not be included in DEEPFISHMAN data archive.

11.2. Fishery-independent survey data

Survey data are expected to be primarily provided under DATRAS or MEDITS data. Other format which is clearly documented is fully acceptable, both the data and format description should be provided as metadata.

Nevertheless, for use of survey data in WP6 biodiversity, survey data may also be provided as the following four files format:

Strata file
- survey name (the same for all records in all files)
- strata code
- strata area
If the survey is not stratified, put one single line in the strata file, one single stratum (e.g. stratum1). If strata surface are not known put 1 for all strata.

Stations file
Data for every haul in the time series
Swept area (NA if unknown), possible to put 1 for all tows if unknown for all
latitude et longitude (preferably in decimal degrees), mean depth

Survey name	Year	Haul number	Strata	Swept area	Latitude	Longitude	Depth

Catch file
Catch by haul by species in number and weight
For species, codes or full scientific names may be used. If codes are use a table of codes and scientific names should be provided.
If some numbers or weights are missing (i.e. for some survey people may have only numbers and no weight), the best is to estimate weights according to mean weight of the species or make any other relevant estimate.
Number and weight in this file are the total number and the total weight by species and by tow. If there are subsamples (i.e. 1000 kg of blue whiting were caught and only 10 kg were counted, weighted measured) we need here the number raised to 1000 kg.

Survey	Year	Haul	Species	Number	Weight

Length file
Number per length class for all measured species.
Again, these are the total number and total weight by species, length class and tow.
All data are not mandatory, there may be only Survey, year, haul, species, Number. In that case other variables may not be provided or set to NA (Not Available)
Weights (by size class) and age are not mandatory.
It is recommended to code sexes as: I Unsexed/immature; F females and M males.

Survey	Year	Haul	Species	Sex	Maturity	Length	Number	Weight	Age

11.3. Biological data

Biological data and life history data should be provided is the case study report. Data file of any format with relevant metadata may be included in Case Study data.

11.4. Ecosystem, biodiversity and VME data

There is no standard format for such data. Data file of geographical locations of VMEs, catch of PETs species, faunal list etc… of any format should be provided to the project.

12. Data policy

The attention of partners is drawn to the following section of the Grant Agreement, Annex II, Part C (Intellectual Property Rights, Use and Dissemination), Article II.30, page 28 of FP7 Grant Agreement - Annex II – General Conditions, Version 2, 02/10/2008

1. Each *beneficiary* shall ensure that the *foreground* of which it has ownership is disseminated as swiftly as possible. If it fails to do so, the *Commission* may disseminate that *foreground*.

2. Dissemination activities shall be compatible with the protection of intellectual property rights, confidentiality obligations and the legitimate interests of the owner(s) of the *foreground*.
In *projects* funded by the European Atomic Energy Community, *dissemination* activities shall also be compatible with the defence interests of the Member States within the meaning of Article 24 of the Treaty establishing the European Atomic Energy Community.

3. At least 45 days prior notice of any *dissemination* activity shall be given to the other *beneficiaries* concerned, including sufficient information concerning the planned *dissemination* activity and the data envisaged to be disseminated.
Following notification, any of those *beneficiaries* may object within 30 days of the notification to the envisaged *dissemination* activity if it considers that its legitimate interests in relation to its *foreground* or *background* could suffer disproportionately great harm. In such cases, the *dissemination* activity may not take place unless appropriate steps are taken to safeguard these legitimate interests.
The *beneficiaries* may agree in writing on different time-limits to those set out in this paragraph, which may include a deadline for determining the appropriate steps to be taken.

4. All publications or any other *dissemination* relating to *foreground* shall include the following statement to indicate that said *foreground* was generated with the assistance of financial support from the *Community*:
The research leading to these results has received funding from the European Community's Seventh Framework Programme (*FP7/2007-2013*) under *grant agreement* n° 227390, project Deepfishman.

Any *dissemination* activity shall be reported in the plan for the *use* and *dissemination* of *foreground*, including sufficient details/references to enable the *Commission* to trace the activity. With regard to scientific publications relating to *foreground* published before or after the final report, such details/references and an abstract of the publication must be provided to the *Commission* at the latest two months following publication. Furthermore, an electronic copy of the published version or the final manuscript accepted for publication shall also be provided to the *Commission* at the same time for the purpose set out in Article II.12.2 if this does not infringe any rights of third parties.

13. Data archives

13.1.1. Stock assessment data

Stock assessment data for all stocks assessed in all case study of the project were uploaded to the project web site (http://www.ifremer/fr/deepfishman), section Zone for partners, Deepfishman data archive, Assessment data.

13.1.2. Other data

Other data by case study were uploaded on the DEEPFISHMAN wiki (http://deepfishman.hafro.is/) section partners.

References

ICES, 2009. Definition of Standard Data-Exchange Format for Sampling, Landings, and Effort Data from Commercial Fisheries ICES Cooperative Research Report, N° 296, July 2009. 48pp.

Annex 1:
Meta data CPR files CS2

File: cpr_FR_OTB_month.csv
CPR stands for Catch Per Rectangle
Data in the file are for single bottom otter trawl only.

This file was created from EU logbook data extracted from Ifremer effort and landings statistics. Only EU logbooks information was used to create this file. The disaggregated data is described below, it cannot be made available to the project as it includes vessels identity. It was aggregated by rectangle and month (i.e. catch and effort in a given rectangle and month were cumulated) this is believe to conserve properly the useful information for fleet strategy or stock dynamic analysis.

Disaggregated data

Disaggregated data are logbook lines. A vessel should report a new logbook line time every day or every time it change statistical rectangle or fishing gear. Each line include catch in weight per species and fishing effort (time fishing) in a given rectangle/day with a fishing gear. Each logbook entry is called a fishing subtrip as especially for offshore vessels fishing trips last for several days and several statistical rectangles may be visited.
In terms of time, there are two effort variables h.fish and h.sea in the file.
h. fish is the number of fishing hours in the subtrip.
h.sea is the number of hours at sea. It may include steaming time. Nevertheless, the time series of h.sea has been considered more reliable than h.fish [this aspect needs some revisiting, please check for consistency if h.fish is to be used].
!! This disaggregated data are individual and nominative and cannot be distributed out of Ifremer !!. In other word this disaggregated data is not available to DEEPFISHMAN, the following data file was created for DEEPFISHMAN

Aggregated data: file cpr_FR_OTB_month.csv

Based on the file above data were aggregated by rectangle/year/month. Fishing time variables are less meaningful here as there are vessels with high and low power. Then, fishing time was multiplied by fishing power and then aggregated.

Columns names in the files .
Three upper case species code are FAO codes. The reporting of separated deepseea sharks species have not been reliable. In early years, GUQ (*Centrophorus squamosus*) and CYO (*Centroscymnus coelolepis*) were reported combined as DGX. The code DGX main include smaller quantities of other deep-water shark species. Therefore, the best options might be to merged DGX, CFB, GUQ and CYO as aggregated deepsea shark landings. Nevertheless, reported landings of individual species were kept to allow for exploration of trends at species level.

Table 1. Detail of variables in disaggregated data (not distributed)

Variable	Detail
vessel	
year	
month	
day	
rect	Statistical rectangle
h.fish	Hours fishing
h.sea	Hours at sea
tcatch	total catch (all species included, kg)
BLI	Blue ling (*Molva dypterygia*)
GFB	Greater Forkbear (*Phycis blennoides*)
RNG	Roundnose grenadier (*Coryphaneoides rupestris*)
ORY	Orange roughy (*Hoplostethus atlanticus*)
BSF	Black scabbardfish (*Aphanopus carbo*)
DGX	Siki (GUQ and CYO) combined
CFB	Black dog fish (*Centroscyllium fabricii*)
GUQ	Leafscale gulper shark (*Centrophorus squamosus*)
CYO	Portuguese dogfish (*Centroscymnus coelolepis*)
power	Fishing power (kw)
kw.f.h	Aggregated h.fish*power
kw.s.h	Aggregated h.sea*power

All species catch are in kg.

Table 2. Detail of variables in the file cpr_FR_OTB_month.csv

Variable	Detail
year	
month	
rect	Statistical rectangle
h.fish	Hours fishing
h.sea	Hours at sea
tcatch	total catch (all species included, kg)
BLI	Blue ling (*Molva dypterygia*)
GFB	Greater Forkbear (*Phycis blennoides*)
RNG	Roundnose grenadier (*Coryphaneoides rupestris*)
ORY	Orange roughy (*Hoplostethus atlanticus*)
BSF	Black scabbardfish (*Aphanopus carbo*)
DGX	Siki (GUQ and CYO) combined
CFB	Black dog fish (*Centroscyllium fabricii*)
GUQ	Leafscale gulper shark (*Centrophorus squamosus*)
CYO	Portuguese dogfish (*Centroscymnus coelolepis*)
kw.f.h	Aggregated h.fish*power by vessel in the month-rect
kw.s.h	Aggregated h.sea*powerr by vessel in the month-rect

All species catch are in kg.

Annex 2:
Metadata for file DWS_fleet_FRA_2001_09.txt

The file includes French vessels that were licensed to prosecute deep-water fishing from 2003 to 2009 and French vessels that landed deep-water species in 2001 and 2002 (i.e. before the implementation of the licensing scheme).
Deep-water licenses where issued in application of council regulation (EC) No 2347/2002 of 16 December 2002 establishing specific access requirements and associated conditions applicable to fishing for deepsea stocks. The regulation entered into force on 01/01/2003.

Variables in the file

Column	Variable name	Description
1	year	Year for which the license applies. Vessels that were licensed for several years appears as many time in the file
2	vessel_id	Vessel Identifier
3	startDate	Date of start validity of license in a given year (1)
4	endDate	Date of end validity of license in a given year (1)
5-29		Vessels data from the Common Fleet Register (CFR) (2)

(1) Licenses may not be valid all years. A new vessel may enter the fleet at any time in the year and an old vessel may be decommissioned. Also, if a vessel is sent to the shipyard for a time, then it is no longer in records of active vessels during this time and its license is suspended. In 2003-2006, licenses have been attributed for full year. Only from 2007, licenses may apply to part of the year only in application of the rules above.
(2) see CFR (http://ec.europa.eu/fisheries/fleet/index.cfm) for detail on variables