PROTOCOLS FOR COLLATING AND STANDARDISING ARCHIVED INDUSTRY DERIVED BIOLOGICAL SPECIES DATA

Report No: 14/J/3/03/2565/1616

Prepared for:
David Cotton and Claire Postlethwaite
MEDIN c/o BODC
Joseph Proudman Building
6 Brownlow Street
Liverpool, L5 3DA

Client Reference Number:

2	Final	C. Edwards	D. Powell	M. Gibson	04/06/2014
1	Draft	C. Edwards	D. Powell	M. Gibson	25/02/2014
Rev.	**Description**	**Prepared**	**Checked**	**Approved**	**Date**

SIGN OFF PAGE

Document Release and Authorisation Record

Fugro EMU Document No.	14/J/3/03/2565/1616		
Issue Number	1		
Hard Copy	0		
Issued To	Claire Postlethwaite		
Date Issued	5 June 2014		
This Version Authorised By	**Name**	**Date**	**Signature**
Author	Dr Cassian Edwards		
Technical Checker	Dr Dawn Powell		
QA Proof Reader	Justine Saunders		JSaunders
Project Manager	Justine Saunders		JSaunders
Group Manager/Director	Mark Gibson		

Contact Details
Fugro EMU Limited
Trafalgar Wharf (Unit 16)
Hamilton Road
Portchester
Portsmouth
Hampshire
PO6 4PX
United Kingdom
Tel: +44 (0)2392 205 500
Fax: +44 (0)2392 205 555
www.fugroemu.com

CONTENTS

1. INTRODUCTION

There is a current drive to facilitate greater access for data users to industry-derived marine environmental data, alongside already publically available data, to reduce the burden of data collection and subsequent costs. Greater availability of historic environmental data to support, for example, environmental assessments, baseline reporting and habitat mapping projects, provides greater data resolution and increased confidence levels in assessments, which in turn facilitates the Regulatory decision making and licensing process.

There are standards and protocols relating specifically to metadata provision, data quality assurance, data format and taxonomic nomenclature that exist, such as MEDIN (Marine Environmental Data and Information Network), NMBAQC (the National Marine Biological Analytical Quality Control scheme), UNICORN and WoRMS (World Register of Marine Species, which encompasses the geographic subset Marine Species of the British Isles and Adjacent Seas - MSBIAS) to try and reduce inconsistencies between environmental survey designs and data reporting. However, despite the efforts of MEDIN and others to try and ensure survey, reporting and metadata consistency, it must be appreciated that obligations by industrial sectors to disseminate their environmental metadata to any existing data portals has not been consistently enforced as a statutory requirement by the Regulators to each sector. Because it has not been a compulsory requirement for much of the time, and is still under Regulator consideration, the extent of the information provided by each industrial sector conducting data gathering exercises in the marine environment are not always of the same breadth and quality. The age of the dataset also has a bearing on how the information is disseminated, if at all. It may be that some metadata compilers have provided a comprehensive amount of information making it easier to locate datasets of interest, but others may have provided only a minimum amount of details as necessary to conform to the MEDIN standard. This can make it more problematic for an end-user to find an appropriate dataset or know of its existence.

Nevertheless, the marine species and associated environmental data currently provided by portals including MEDIN, The National Biodiversity Network (NBN) gateway and The Crown Estate's Marine Data Exchange (MDE) are valuable commodities and in order for the data to be shared for the purpose of temporal or spatial comparisons then there is a need to make sure that the end-user has a clear understanding of which datasets are fit for comparative purposes. For example, can data obtained by one study be compared with data obtained by another? Likewise, can data recorded by different methods be compared, such as remote video data with data recorded by divers? In answering these questions there exists a requirement to perform checks for inconsistencies between datasets collected using the same or similar methods, before comparisons are undertaken so as to prevent potentially misleading and invalid results from occurring.

The aim of this document is to present a protocol for those using the MEDIN supported Data Archive Centres (DACs) for the collation of datasets containing biological species data, but with consideration of the accompanying supporting data. Although this protocol is primarily designed for MEDIN, the principals apply for data collation from any data repository to assess whether the datasets are suitable for the end-user's requirements, and additionally, for comparative purposes. This protocol outlines the various data prerequisites and assumptions that need to be satisfied when selecting appropriate species datasets. The document also provides the taxonomic reconciliation steps that may need to be conducted before using a species dataset for statistical analyses. To aid the process, a simple flow-chart of the protocol is presented. Note: it is assumed that these tasks will be performed by practicing marine ecologist/taxonomist with expert knowledge of community data interpretation.

2. FINDING SPECIES AND ASSOCIATED DATA AND ASSESSING IT FOR USE

Validation of whether or not species datasets and supporting environmental data are fit for comparative purposes is a three-step process. Firstly, metadata are queried through MEDIN's Data Discovery Portal to find datasets which may be applicable to the data end-user's requirements in relation to the study in question. This stage is termed here as a *Metadata Assessment* (MA). The different metadata elements of the datasets are interrogated and assessed as to the appropriateness of the data. Some metadata records may contain all the information needed to verify that a dataset is fit for the purpose of the end-user's project. However, additional information which is not contained in the metadata may be needed to assess dataset appropriateness.

Instructions provided by MEDIN prompts a metadata compiler to input comprehensive information about a dataset when writing metadata. When performing the MA, it is imperative the end-user reviews the data entry guidelines/instructions for the data entry procedure provided on the portal to know which fields may contain the detail required (i.e. the 'Abstract' or 'Lineage' fields). However, this current protocol accepts that the process of data entry is subjective and possibly performed by a third party who may not understand the importance/nature of the detail requested in the fields. It is possible that the information relevant to the end-user is not provided which makes it hard to find a suitable dataset for the requirements of a project. Consequently, the methodology used to obtain the data, as well as other factors, such as the exact sampling site coordinates, may then need to be screened. If applicable, screening is accomplished by accessing the technical reports and field logs associated with the dataset to verify the site coordinates and sampling and processing methods. This stage is termed here as a *Data Fitness Assessment* (DFA). Once these checks are complete and the data have been confirmed as suitable, the two (or more) datasets then have to pass a *Taxonomic Reconciliation Assessment* (TRA). The TRA is performed to check that there are no discrepancies in the taxonomic nomenclature and recording of the different species which would present as artefacts at the results stage. Only after the TRA has been undertaken, and any taxonomic amendments performed as necessary, should a statistical analysis of the data proceed.

The different stages that are needed to find data, assess data fitness, and then reconcile the taxonomy between the different species datasets are discussed in more detail below.

2.1 Metadata Assessment (MA)

The MA aims to select potential datasets for analysis, based on the metadata elements associated with the dataset. The MA is a logical step in the process of finding data. Queries are made via the MEDIN data discovery portal and the returned results list 'projects' (which could be in the form of datasets and/or surveys) matching the search query terms. The metadata elements of the different projects that are returned are then screened by the end-user to assess if the data available is applicable to the needs of the project. Applicability is partly determined by the extent of the detail recorded in the metadata elements, which is largely dependent on the level of detail provided by the metadata compiler and the various options available for input.

It is acknowledged that MEDIN places a large emphasis on metadata compilers to fill in comprehensive metadata details, where applicable. However, it cannot be assumed that the detail required by an end-user to select an appropriate dataset will be contained within the metadata, even if the relevant detail could be included when the metadata is compiled and submitted to MEDIN. Although many studies provide information in their metadata which is needed to screen potentially suitable datasets, such as the sieve mesh aperture size used to process a sediment sample for macrofauna, many metadata records do not provide this level of detail. In this situation, this protocol will drive the end-user to gain access to the technical reports, including associated field logs and/or information regarding the survey methodology, as discussed in section 2.2.

2.2 Data Fitness Assessment (DFA)

After the MA has been conducted to find potential datasets for analysis, a DFA will need to be undertaken. The purpose of the DFA is to screen potential datasets that may be applicable to the study objectives, to assess the level of comparability and consistency in the survey methodologies, survey equipment, and data processing stages. Specific factors relating to the end-user's project aims and objectives may need to be known to assess whether a dataset is fit for the purpose of answering specific marine ecological questions. These factors include:

- Location of study area and sampling sites:
  - Exact coordinates of survey area boundaries and sampling sites may be required. These coordinates may need to be plotted in a GIS to confirm validity of site selection, and the exact locations from where the data have been obtained.
  - It is vital that the Coordinate Reference System is taken into account when comparing locations. No assumptions should be made that all coordinates are in a standard format such as WGS84. Location checking should be performed by someone with knowledge of global coordinate systems.
- History of the study area:
  - Knowledge of any past site events. For example, whether or not the area was in an impact zone. Whether the study was a characterisation study or an impact related study would hopefully be clear from the metadata, but GIS layers may need to be visualised to see exact site locations (preferably overlaying locations on the latest Admiralty Chart).
- Type of survey:
  - Grab survey;
  - Trawl – 2m beam trawl, 4m beam trawl, etc.;
  - Drop-down video, diver survey or remote sensing.
- Exact specifics of sampling regime and survey equipment:
  - Grab type, trawl type;
  - Area, depth and volume of sampling unit;
  - Mesh size, cod-end size, trawl speed, tow length, duration, etc.;
  - Were there replicate samples?

- Target and extraneous organisms in the dataset:
  - Macrofauna – does the dataset contain infauna or epifauna or both?
  - Meiofauna;
  - Fish or shellfish;
  - Juveniles / larval stages.
- Habitat characteristics of the survey area
  - Are there any environmental data, including sediment type, depth, salinity, and temperature associated with the species datasets? This information can be very important if like-for-like comparisons between different survey areas are to be made.
- Processing methodology:
  - How were the organisms extracted from the sediment?
  - What sieve aperture size was used to extract organisms from sediment? This information is extremely important since valid comparisons cannot be reached by comparing macrofauna datasets obtained via processing with different sieve sizes.
  - Has sub-sampling been undertaken?
  - Is the extraction efficiency known?
- Data history:
  - Are the data in an original raw data format or have they been taxonomically rationalised?
- Data format:
  - What format are the data in? Absolute abundance, percent abundance, biomass, percent coverage, SACFOR scale, presence/absence?
- What area / volume units are the data presented in?
  - 0.1 m^2, 0.2 m^2, 1 m^2?
  - Can different units be standardised?

Note: the metadata contained in the MEDIN portal is "discovery" metadata, i.e. it helps others discover that a dataset exists. It is not intended that the level of detail mentioned in section 2.2 be included in any of the fields in the portal. However, if this level of detail has been provided, it is likely to in the ‘Abstract’ or ‘Lineage’ fields. The level of detail that is required to compare datasets, such as sieve mesh size etc, will be stored with the dataset. This is known as "use metadata" and may already be available as a spreadsheet rather than hidden in a report, especially if the data is available at one of the MEDIN Data Archive Centres.

A rapid screening process will reveal if these factors are detailed in the metadata records. It is possible that the field logs and methodologies of the different datasets returned by the MEDIN metadata search will need to be interrogated. This is to ensure that data obtained for the purpose of statistical analysis or otherwise are valid for answering the questions being asked and the hypotheses being tested. The validity of data is determined by scrutiny of metadata elements between datasets to select a number of potentially suitable studies, and then by interrogation of the finer detail contained in the associated Technical Reports (or field logs and sampling and processing methodologies), where necessary. If statistical analysis of the dataset is planned, performing diligent checks on these finer details will prevent artefacts from arising at the results stage due to differences in survey methods and sample processing, such as differences in grab or sieve aperture size between datasets, respectively.

Hypothetical example

An academic research study or a developer might want to perform a statistical analysis of change in benthic macroinfaunal abundances at a southern North Sea aggregate extraction licence area over time. To make valid statistical comparisons, which are free from methodological artefacts, there should be consistency in the methods used to obtain the data. The data used in the analysis would ideally have been obtained using the same type and size of benthic sampling grab (although different grab types may be applicable, with caveats), from one time period to the next, in order to make valid conclusions regarding whether or not there have been any temporal changes in benthic infaunal abundances. A metadata search of the data discovery portal for ‘Southern North Sea’ and ‘macrofauna’ brings up a number of results. However, these results may not necessarily state the exact sampling methodology and the sieve aperture size (e.g. 1 mm or 0.5 mm) used to extract the macrofauna from the sediment. This knowledge of the size of the sieve aperture is imperative to selecting a suitable dataset for comparative purposes, and drawing valid statistical comparisons between datasets. This information may therefore have to be obtained by approaching the Responsible Party for the dataset, or the online repository where this information may be held. Only once it is verified that the macroinfaunal data have been obtained by processing sediment through the same sieve aperture size can the data be deemed as statistically comparable, subject to a Taxonomic Reconciliation (section 2.3). The same applies to trawl datasets where different sized cod-ends may have been used, different core sizes/depths in meiofaunal studies, etc. If this information is not detailed in the metadata, further interrogation of Technical Reports and associated datasets may be necessary to select appropriate data for analysis.

2.3 Taxonomic Reconciliation Assessment (TRA)

The next part of the protocol may involve a TRA. The purpose of the reconciliation is to standardise the taxonomy in two or more datasets if statistical comparisons are to be made. If no statistical analysis is planned then raw data may not need to be reconciled. Statistical comparisons between historical data and more recent data may not be valid due to fundamental differences in the way in which the different sets of data have been obtained and the temporal separation may be considered too great. Nevertheless, historic data **should not be discounted altogether** as it is a snapshot of a community in time. These data can provide important information about the varying assemblages and biotopes that an area has supported in the past and can be treated as an independent dataset for comparison at the community/biotope level.

The aim of the TRA is to prevent artefacts from arising at the **statistical analysis results stage** due to inadequate taxonomic reconciliation across the different datasets. This is a fundamental stage if datasets are to be statistically compared in a software package such as PRIMER. Although datasets may have passed the MA and DFA stages, a TRA will still need to be performed since discrepancies can arise due to sourcing of data from different surveys, regions, time periods and also differences in laboratory outputs (i.e. taxonomic resolution). These discrepancies manifest as organism names changes over time (e.g. *Pomatoceros triqueter* has changed to *Spirobranchus triqueter*), the reporting of individuals of the same species to different levels of taxonomic hierarchical classification (often due to variable levels of identification experience at different laboratories), inclusion of additional taxonomic qualifiers such as sp. or spp., as well as the inclusion or exclusion of juvenile individuals (such as *Abra* juv.). In addition, data may be recorded in several different formats, including absolute abundance, % coverage, presence/absence, or recorded on the MNCR SAFCORN[1] scale, amongst other formats.

Examples will be presented in the following paragraphs utilising two hypothetical datasets, termed ‘dataset 1’ and ‘dataset 2’. These datasets could pertain to data collected from the same site *if temporal comparisons were to be made*, or from different sites *if spatial comparisons were to be made.* These hypothetical datasets will be used to illustrate commonly encountered data reconciliation problems, which the protocol presented in this report endeavours to rectify. The term ‘record’ hereafter will be used to imply a single row (i.e. taxon) in a two-way *species* by *site* (or *time*, etc.) matrix, where *species* are rows and *sites* (or *times*) are columns.

The TRA stages should be performed by someone proficient in macrofaunal taxonomy and statistical analyses to ensure a sensible and realistic outcome.

2.3.1 TRA Stage 1 – Nomenclature Discrepancies

Nomenclature discrepancies manifest as taxonomic name changes over time, spelling mistakes, and inclusion of taxonomic qualifiers. A taxonomic qualifier is an additional identifier to a taxonomic record, such as ‘sp.’ or ‘spp.’, ‘male’ or ‘female’, ‘juv.’, ‘agg.’, ‘indet.’, etc. The inclusion of taxonomic qualifiers in a dataset may be of importance to the aims and objectives of the study; however qualifiers may also be extraneous for comparative purposes and may need to be removed.

As a first step, the different species datasets should be combined into a single spreadsheet and all qualifiers removed. The list of species contained in this spreadsheet should then be checked for spelling mistakes and changes in taxonomic nomenclature via the WoRMS taxon match tool (http://www.marinespecies.org/aphia.php?p=match), or the geographical subset tools (e.g. MSBIAS) as geographically appropriate. The taxon matching tools spot inaccurate and out-of-date nomenclature and provide a correct species list in accordance with currently accepted names. Taxon matching is a crucial step particularly where temporal comparisons are being undertaken between datasets. As an example, consider a comparison between dataset 1 and dataset 2. The keel worm may have been recorded as *Pomatoceros triqueter* in dataset 1, but as *Spirobranchus triqueter* (its currently accepted WoRMS compliant name) in dataset 2. If a temporal statistical analysis were to proceed without having first amended the name of this species to *Spirobranchus triqueter* in dataset 1, the results would likely suggest that there had been a decline in *Pomatoceros triqueter* but an increase in *S. triqueter* between the two datasets. This is obviously an erroneous and artefactual conclusion to reach, since all individuals are from the same species, albeit labelled using the old taxonomy in dataset 1 and the current taxonomy in dataset 2.

Once the species list is free from spelling mistakes and taxonomic errors, it should be reassigned to the dataset. Since statistical packages such as Plymouth Routines in Multivariate Ecological Research

---

[1] S = Superabundant, A = Abundant, C = Common, F = Frequent, O = Occasional, R = Rare, N = Not Recorded

(PRIMER) treat duplicate species records as two different species, any duplicate records will need to be located and subsumed into one record, as appropriate.

A decision will now need to be made regarding the treatment of taxonomic qualifiers. Records which initially had taxonomic qualifiers should be assessed and, where appropriate, have their qualifier replaced. The replacement of qualifiers should be governed by the aims and objectives of the study, and whether or not the qualifiers are important to the questions being asked. If the qualifiers are not important then they should not be replaced. For example, a study looking at crustacean population structure may need to know whether individuals of the edible crab *Cancer pagurus* were male or female. In this case records with a gender qualifier should remain. However, this level of identification detail may not be appropriate for a study on macroinvertebrate community structure and therefore the two records may need to be subsumed into a single record labelled *Cancer pagurus*. Subsumption should be conducted with care since data in different formats, i.e. absolute abundance, % abundance, presence / absence etc., may need to be subsumed in different ways.

Recently, WoRMS has updated its capability to cope with qualifiers such as “sp.” and “spp.” and therefore these do not need to be removed from the dataset before matching. However, the use of these qualifiers imposes specific information about the data recorded; "sp." is used when the actual specific name cannot or need not be specified, whereas the abbreviation "spp." (plural) indicates "several species". As with other qualifiers, a decision needs to be made as to whether or not records containing these qualifiers are to remain *as is* within the dataset, or whether individuals denoted as *Xyz* sp. or *Xyz* spp. can be grouped into a single record containing the genus *Xyz* without any additional qualifier. This decision will be project specific and will depend on the questions being asked and any project-specific requirements.

2.3.2 TRA Stage 2 – Rationalising the Level of Taxonomic Identification

The next stage is to identify records where individuals of a single species have been identified at different taxonomic levels in the different datasets. For example, individuals of the species *Spirobranchus triqueter* may have been previously identified as *Spirobranchus* sp. in dataset 1 but as *Spirobranchus triqueter* in dataset 2. Although both records refer to the same species, all records will have to be raised to *Spirobranchus* sp., the lowest common denominator.

The publication of new or updated identification keys also poses data reconciliation problems. Recently, the publication of updated keys has facilitated increased confidence in the identification of several species in the *Syllis* genus to species level. Therefore, when comparing datasets where the species have been identified to genus only with datasets which have identified individuals in the genus to species level, a decision needs to be made whether to group the abundance of all individual species to genus level and lose some data resolution.

2.3.3 TRA Stage 3 – Dealing with Juveniles

Depending on the purpose of the analysis, records of taxa with reference to juvenile stages (e.g. *Abra* juv., Mytilidae juv., Cirripedia juv.) should be considered with caution, and a decision will need to be made regarding the inclusion or exclusion of these records. This decision should be based on the purpose and reasons for the analysis. If the purpose of the analysis is to look at differences in juvenile and adult age classes, for example, then the inclusion of juvenile stages should be left *as is*. However, if the analysis is for regional monitoring purposes then the relevant OSPAR guidance document will need to be considered. OSPAR (2004) states that “Newly settled juveniles of benthic species may at times dominate the macro-fauna numerically, but due to heavy natural post-settlement mortality, such dominance must in many cases be regarded as transitory and not as an indicator of the prevailing bottom conditions. Should juveniles appear among the ten most dominant organisms in the data set, the statistical analysis should be conducted both with and without these in order to evaluate their importance.”

2.3.4 TRA Stage 4 - Removal of Recorded Taxa that are out of Context in Relation to Study Aims

The final stage is to remove certain taxa that may have been recorded but are out of context in relation to the aims of the study, such as meiofauna. Meiofaunal organisms consist of a taxonomically diverse group of metazoans and protozoans which are generally smaller than macrofauna (e.g. <1 mm) yet bigger than the nanofauna (e.g. bacteria, microalgae and most protozoans) (Coull and Bell, 1979). Some larger individuals of meiofauna, such as nematodes will, however, be retained by a sieve with a 1 mm aperture, the same size which is typically used to separate macrofauna from sediment samples (e.g. Ware and Kenny, 2011). This results in a high degree of under-recording of the meiofaunal community and thus makes such data irrelevant.

Records of other taxa which may need to be removed include those which have been identified to a low level of taxonomic differentiation, such as individuals identified as “Crustacea” or “Mollusca”, for example.

It should be noted that certain fish species which are recorded in a grab dataset may need to remain within the dataset, particularly if it is being assessed for biotope classification. For example, sandeels *Ammodytes* spp. are indicative of the ‘infralittoral sand’ biotope, and as such should not be removed from a benthic grab dataset which is to be biotoped.

3. SUMMARY

The itemised points below represent a summary of the series of successive steps recommended and detailed above.

i. Search MEDIN Data Discovery Portal for potentially applicable biological species dataset(s) and supporting environmental dataset(s). Guidance on searching for UK marine data is provided at: http://www.oceannet.org/extlink/http%3A//bit.ly/Hwo6Op

ii. Assess each dataset for applicability to your project by examining the metadata elements associated with each dataset – a *Metadata Assessment* (MA). (Always refer back to the specific user guidelines on data entry to ascertain which user fields should contain the necessary metadata).

iii. Query the metadata for a *Data Fitness Assessment* (DFA). If the metadata provides all the necessary information to help you decide if the data are suitable for your needs, either contact the Responsible Party to request access to the data or download the data from an online repository, as appropriate.

iv. If the metadata do not provide all the necessary information to help you decide if the data are suitable for your needs, then the Technical Report (and associated field logs and survey and processing methodology) may need to be screened to verify that the data contained in the datasets have been obtained in a manner consistent with, or appropriate for, your needs. Sampling locations from the field logs may need to be plotted using a GIS package to confirm exact site positions that data have been obtained from.

v. Once an appropriate dataset has been obtained, if statistical analyses are being applied, conduct a *Taxonomic Reconciliation Assessment* (TRA) of all the datasets to ensure they are comparable. Taxonomic differences will manifest themselves as statistical artefacts if this process is not performed.

vi. Only once the dataset has passed the MA, the DFA, and the TRA stages can it be deemed fit for the statistical analysis stage.

vii. If the data are not being manipulated statistically, they do not need rigorous rationalisation, but the end-user may need to consider historic names changes nevertheless. Historic datasets have intrinsic value regardless of the data quality; key characterising species should still be obvious and they provide a snapshot in time. However, the way they have been obtained and dealt with must guide the end-user to their appropriate treatment and reporting.

4. REFERENCES

Coull, B.C. and Bell, S.S., (1979) Perspectives of marine meiofaunal ecology. In: Livingston, R.J. ed. *Ecological Processes in Coastal and Marine Systems.* Plenum Publishing Corporation, New York. pp. 189-216.

OSPAR (2004) OSPAR Guidelines for Monitoring the Environmental Impact of Offshore Oil and Gas Activities. Reference number: 2004-11E.

Ware, S.J. and Kenny, A.J., (2011) Guidelines for the Conduct of Benthic Studies at Marine Aggregate Extraction Sites (2nd Edition). Marine Aggregate Levy Sustainability Fund, 80 pp.