



**THE MARINE BIOLOGICAL ASSOCIATION OF
THE UNITED KINGDOM**
Data Archive for Seabed Species and Habitats



The Data Archive for Seabed Species and Habitats (DASSH)
Quality Assurance (QA) procedures.

1. Data accessions.

The internal DASSH accession system has been developed to monitor data acquisition, archive, input and quality assurance (QA), which allows each dataset to be tracked from initial contact with the data provider to the completion of the archive process and data dissemination. The steps in the data acquisition process are outlined in Figure 1. Each stage is recorded in the accessions system via a web interface, along with the name of the person responsible for undertaking it (see Appendix 1). Log files of data requests are linked to the accessions system so that all standard downloads and requests for use of each dataset are recorded.

The accessions information, dataset requests and downloads of data will be made available online to data providers (via a unique provider login) so that they can monitor the progress of datasets in DASSH and the interest shown in their datasets.

In an ideal situation, the permission for a dataset would arrive first and the metadata and data would follow together. However, it is not always possible to collect data in this order, as data providers may email data before sending permissions etc. The permissions, metadata and data paths are therefore designed to be flexible. The data will only be disseminated once all three data accession criteria are fulfilled.



Figure 2 - Flow chart of accessions steps.





Date Original material
returned, (if applicable)

2. Archive of data

Raw datasets are archived on the DASSH archive server as soon as they are provided (if digital) and as soon as practically possible if provided in a non-digital format. As the dataset is progressed through DASSH, metadata documents, processed data and QA records are added to the archive. Any physical copies of data or metadata provided are held in a locked fireproof cabinet.

3. Quality assurance of data and metadata.

Raw data

Raw data is initially checked against the metadata provided to ensure that the data provided matches the description of the data given. Data points are plotted in GIS to ensure that co-ordinates fit in the bounding box and match the reported extent of the dataset, i.e. does a Thames Estuary survey produce data points that map in the Thames Estuary. Records of data collection and other QA procedures applied by the data provider are recorded in the lineage element of the metadata.

Species records are checked against the NBN / NHM Species Dictionary on entry to Marine Recorder and using in-house macros to check Excel spreadsheet species information. Biotope data is checked against the relevant classification (i.e. Connor *et al.*, 1997; 2004). Any unusual or odd records are referred to the data provider and, if necessary, to local expert marine biologists for confirmation. Unusual records include new species sightings, species outside their recorded distribution range and/or in unexpected habitats. Where data has been collected without quality assurance processes, for example volunteer records, then unusual records would need confirmation via either a specimen or photograph which can be referred to a taxonomic expert where required. For historical records where such rigorous confirmation may not be possible then the record will be marked as uncertain.

The raw data is archived in its original form. Any amendments to the datasets prior to dissemination, e.g. typographical errors, species name changes, etc are documented.

Metadata

Manual quality assurance

Initial QA of metadata is undertaken by checking the metadata information for each field against the source document(s) (e.g. data files and reports). Metadata cannot be entered

Formatted: Bullets and
Numbering

Formatted: Bullets and
Numbering



and QA'd by the same person. If any updates are made the person reviewing the metadata then the date of metadata update is recorded. Where possible, final metadata is referred to the data provider for an additional check to ensure accuracy.

DASSH metadata QA procedure is based on the 'Metadata Guidelines for Geospatial Datasets in the UK' guidelines produced by the Department for Communities and Local Government and published by GI-Gateway (Figure 1.)

Validation against the Marine Data and Information Partnership (MDIP) Metadata XML Schema

Once manually checked, the metadata is converted into an eXtensible Markup Language (XML) document based on the MDIP Metadata Schema. These XML documents must validate against the schema before they can be harvested by the NERC Data Grid software. XML validation is carried out automatically using XMLSPY 2007 software with manual corrections if any document fails validation.

Processed survey data

DASSH routinely plots the survey bounding box and the survey points, using in-house GIS tools, to check the dataset.

Marine Recorder

Where data has been processed in Marine Recorder, the JNCC validation procedures are used to QA datasets (see <http://www.esdm.co.uk/MarineRecorder/index.html> for full documentation relating to the validation procedure). Completed validation forms are stored with the archived data. Data is not entered and QA'd by the same person.

Guidelines for the number of survey events¹ to QA are outlined in Table 1 and some flexibility is allowed to allow for differences in QA of data inputted directly (e.g. from paper reports) and data imported electronically in which data copying errors are less likely. If errors are found in the QA sample the dataset the entire dataset is subject to further QA.

¹ Survey Event - A single geographic location within a survey e.g. a sampling station. Survey events are also temporally distinct so the same location surveyed at different times with in a year would constitute different survey events.

Formatted: Bullets and Numbering



Table 1 - Guidelines for number of survey events per dataset to QA.

Number of survey events	Number of survey events to QA
1-30	50-100%
31-60	25-50%
61-100	20-25%
101+	10-20%

DASSH datasets sent to the National Biodiversity Network

Datasets that are sent to the National Biodiversity Network (NBN) are subject to additional validation. DASSH uses the NBN Validator software to QA records sent to the NBN from Marine Recorder and ensure that data meets NBN requirements.

Dealing with changes in taxonomy.

Taxonomy continues to change, as new research evidence comes to light. In some datasets the taxonomic names provided are no longer recognised and these names are cross referenced with the NBN/NHM Species Directory and European Register of Marine Species (ERMS) database and updated where necessary. In such cases the original taxon and the taxon that it has been replaced with will be included in the lineage element of the metadata for reference purposes.

Archive and quality assurance of image and video data

Digital images

Digital images are archived in the format they arrive. A second processed image (put into a standard format and renamed) will also be archived. Raw digital video will be compressed using H.264 Encoding (the Industry standard for lossless compression) and will be archived in this compressed form only. It is assumed that the digital image or video is supplied in a colour corrected format. Any processing undertaken by the data provider will be recorded in the lineage element of the metadata. Manual checks to ensure the image fits the description given, and that it links up correctly with any survey data are also part of the QA process.

Non-digital still images

Although digital images are preferred, DASSH may accept on a case by case basis images in



**THE MARINE BIOLOGICAL ASSOCIATION OF
THE UNITED KINGDOM**
Data Archive for Seabed Species and Habitats



print, interneg, negative or slide formats. All images will be scanned at 1200dpi on either a NIKON slide scanner or other high quality flatbed scanner. Digitised images will be colour corrected to the original using a high spec monitor that is colour-calibrated on a monthly basis. Once digitised, images will be dealt with as standard digital images.

VHS video

VHS video will be digitized and stored as a compressed file. Video will be compressed using H.264 Encoding,. Once digitized, video will be dealt with as standard digital video.

Quality Assurance of optimised digital media for the website.

Batch processing of images at Web resolution should minimise sources of error. However, DASSH manually checks 30% of images produced to ensure that the:

- filenames are correct;
- the quality of the images are adequately retained, and
- the correct descriptions are linked to the images.

4. Errors

As well as a QA procedure for data DASSH also has an error reporting procedure to manage and deal with any incidences where erroneous data is not picked up by the QA procedure. DASSH aims to respond to any enquiries about errors within 10 working days. Reported errors are logged in the accessions systems along with any steps necessary for their remediation.

Initial checking of suspected errors involves a recheck of processed data against the raw data for copying error. If the error is identified it is corrected. If a suspected error is contained within the raw data it is flagged with the data provider for comment.