THE MARINE BIOLOGICAL ASSOCIATION OF THE UNITED KINGDOM

The Archive for Marine Species and Habitats Data (DASSH)

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

1. Data Accessions

The internal DASSH accessions system has been developed to monitor data acquisition, archive, input, and quality assurance (QA). This 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 (gitlab), along with the name of the person responsible for undertaking it.

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 published and disseminated once all three data accession criteria are fulfilled. DASSH is currently working to develop an interface whereby the accessions information, dataset requests and downloads of data are available online to data providers (via a unique provider login) so that they can monitor the interest and usage information pertaining to their own datasets.

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. Each data set is archived within a folder titled using a unique Resource ID. An instance (issue) is created for the dataset on the accessions system (Gitlab) to track the progress of the dataset through the accessions system. As the dataset is progressed through DASSH, metadata documents, processed data, any additional raw data and QA records are added to the archive folder. 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 (Geographic Information System) software to ensure that the location co-ordinates provided 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. Details of data collection and QA procedures applied by the data provider, including the citation for any associated reports if available, are recorded in the lineage element of the metadata.

Species records are checked against the Marine Species of the British Isles and Adjacent Seas (MSBIAS) or the World Register of Marine Species (WoRMS) directories depending on the location in which data was collected. All taxon names and Aphia IDs are reviewed and updated where required with the accepted details. Biotope data is checked against the relevant vocabulary (i.e., European Nature Information System (EUNIS)). Where there have been changes in taxonomy since the creation of the raw data, the unaccepted names are cross-referenced with the MSBIAS or WoRMS databases and updated where necessary. In such cases the original taxon name will be included in a separate column for reference purposes.

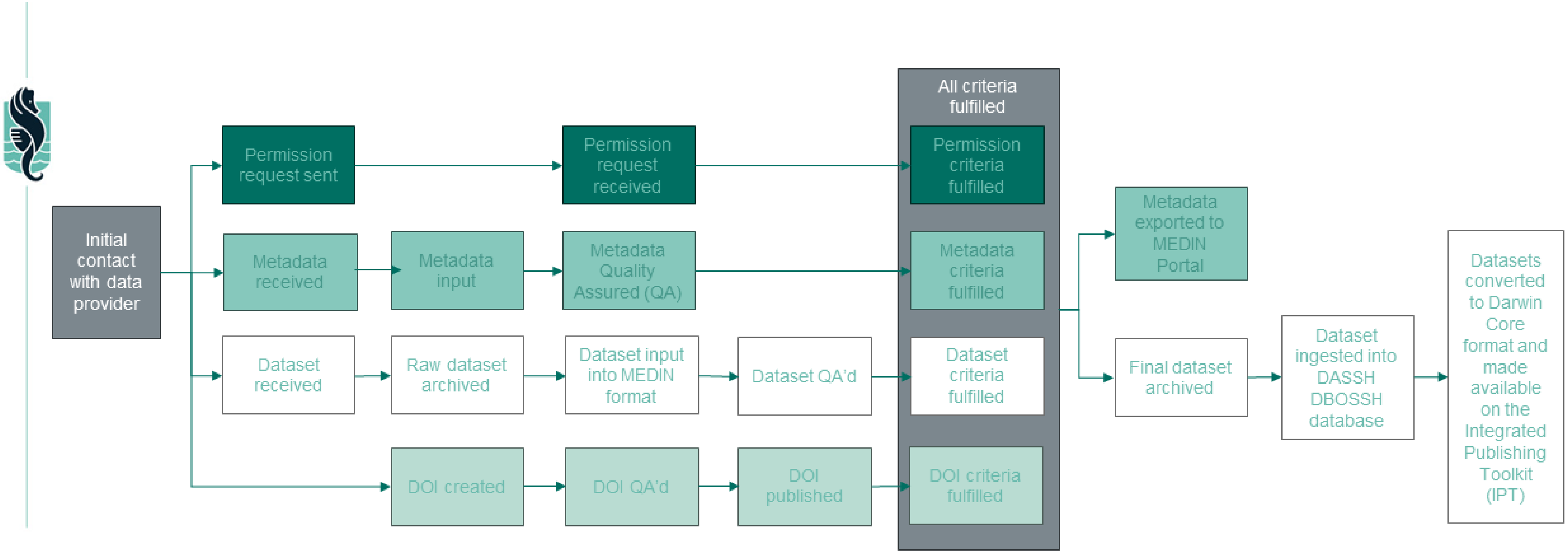
Initial contact with data provider
Permission request sent
Permission request received
Metadata received
Metadata input
Metadata Quality Assured (QA)
Dataset received
Raw dataset archived
Dataset input into MEDIN format
Dataset QA'd
DOI created
DOI QA'd
DOI published
All criteria fulfilled
Permission criteria fulfilled
Metadata criteria fulfilled
Dataset criteria fulfilled
DOI criteria fulfilled
Metadata exported to MEDIN Portal
Final dataset archived
Dataset ingested into DASSH DBOSSH database
Datasets converted to Darwin Core format and made available on the Integrated Publishing Toolkit (IPT)

Figure 1 DASSH Flow chart of accessions steps

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 require confirmation via either a specimen or photograph which can be referred to a taxonomic expert. 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 in subsequent versions.

Where possible the entire dataset is run through the QA process twice by separate members of the data management team. However, with large data sets it is not always possible to QA the entire dataset. Guidelines for the number of survey events[1] and taxon records to QA are outlined in Table 1 with some flexibility allowed for differences in QA of data inputted directly (e.g. from paper reports) and data imported electronically If errors are found during the QA of the dataset, this may be subject to further rounds of QA if necessary.

Following the QA, each dataset is assessed for quality, categorising the data quality as either High, High-Medium, Medium, Medium-Low, Low or Data-deficient as set out in the DASSH Quality Assurance Matrix[2]. The criteria is set out in the ISO 19115 standard for geospatial metadata (ISO, 2006) and uses guidelines set out in Rackham & Walker (2006). This information is added to the metadata to provide insight into the quality of the dataset to potential users[3].

Table 1 - Guidelines for number of survey events and/or taxon records per dataset to QA.

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

Metadata

Manual QA: Metadata is initially created using the MEDIN Metadata Editor. Initial QA of metadata is undertaken by checking the metadata information in each field of the Metadata Editor against the source document(s) (e.g., data files and reports). Metadata cannot be entered and QA'd by the same person. If any updates are made the person reviewing the metadata, then the date of the metadata update is recorded. Where possible, final metadata is referred to the data provider for an additional check to ensure accuracy.

[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 within a year would constitute different survey events.

[2] DASSH (2023): The Archive for Marine Species and Habitats Data (DASSH) Quality Assessment (QA) Criteria. The Archive for Marine Species and Habitats Data (DASSH). (Text). http://doi.org/10.17031/64ccec5c7ba72

[3] Quality Assessment process implemented by DASSH from September 2023

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

Automated QA: The Metadata Editor runs an automated validation against the Marine Data and Information Partnership (MDIP) Metadata XML Schema. 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.

4. Archive and quality assurance of image and video data.

Digital images

Digital images are archived in the format provided. A second image is put into a standard format, renamed, and archived. Raw digital video is compressed using H.264 Encoding (the industry standard for lossless compression) and is archived in this compressed form only.

It is assumed that the digital image or video is supplied in a colour corrected format. Any processing that has been undertaken by the data provider will be recorded in the lineage element of the metadata. A manual check is also undertaken to ensure the image fits the description given, and that it links up correctly with any survey data.

Non-digital still images

Although digital images are preferred, DASSH may accept on a case-by-case basis images in print, 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 monthly. 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 descriptions are correctly linked to the images.

5. Errors

Following publication, any further errors may be identified and reported to DASSH. The team can be contacted using the email dassh.enquiries@mba.ac.uk. The DASSH team aims to respond to any general enquiries, data submissions and 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.