Streamlining and Standardising Pelagic Data Flow to Support the MSFD

Summary

MEDIN funding in 2018 enabled DASSH to develop the infrastructure and tools to support better data management processes for the HBDSEG Pelagics Working Group. This demonstrator project illustrates the benefits of engagement with MEDIN Data Archive Centres, the importance of data standards in the effective collation of biological data and provides a model for how agencies undertaking monitoring activities can leverage maximum value from their data.

Background

In order to ensure the consistent development of tools and services to support the suite of pelagic indicators in the context of the Marine Strategy Framework Directive, DASSH have worked with the UK pelagic habitats indicator lead and Working Group to ensure the underlying data are standardised and suitably described with appropriate metadata.

Following the successful award of the MEDIN funding the DASSH team met with the UK Pelagics Working Group to fully develop the requirements for the work and to ensure concerns over intellectual property and data release were met.

In parallel work began on the ingestion of the existing data from a previous project and re-working of the Python scripts that had previously been used to interrogate the data.

Data Integration

The EcApRHA (Applying an Ecosystem Approach to (sub) Regional Habitat Assessment[1]) project collated a database of ca 42 million plankton species records from 117,316 samples. In addition, the database contains a 'mapping' of each species to a number of physiological and morphological traits, and a mapping of trait combinations to a lifeform category. For more information on the traits and lifeforms used it is recommended to read the OSPAR report (https://www.ospar.org/site/assets/files/36763/1_1_plankton_lifeforms.zip).

The database was originally developed in EcApRHA using a mongoDB non-SQL JSON-document structure. It was decided to convert it into a relational database to improve interoperability with other DASSH maintained data systems. Given the size of the data the resulting database required an iterative optimisation approach in finalising the layout of the new relational tables and the conversion process. Eventually a way was found to convert all 42 million records into the new structure in a reasonable time (~1 hour), and the new relational set-up has proved efficient for the lifeform interrogations needed (see below).

Script Optimisation

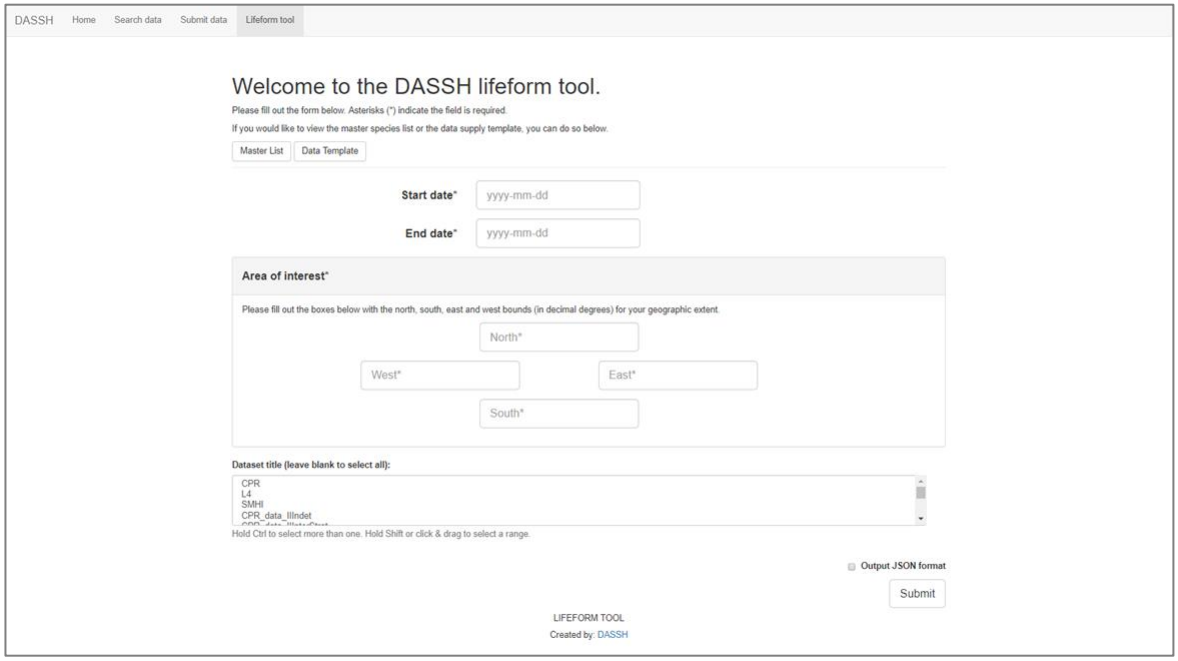
An existing script was developed using the Python language during the EcApRHA project to extract Lifeform information (specifically the average abundance of species in each lifeform category over a given period). This tool needed re-developing as a) it took an unreasonably long time to return results and b) it was designed to interrogate the original mongoDB database rather than a PostgreSQL relational database.

[1] https://www.ospar.org/work-areas/bdc/ecaprha

The finished script allows users to specify a spatial bounding box (based on decimal latitude and longitude) and temporal resolution (date range). Abundances for each lifeform category are provided as a per sample average for each month as well as for the entire time period. Improved script and database efficiency has enabled acceptable search times. For reasonably small search areas (e.g. a range of a few months) it takes just a few moments to return results, for larger timescales (e.g. a decade) the tool takes around five-ten minutes based on the current dataset size. The results are returned either as a JSON string or formatted and printed to screen. Future work will allow the download of data in CSV format for integration with other tools including CefMat.

Front-end Development

A web portal was created and hosted on the DASSH website to interface with the lifeform extraction Python script. The portal is available at 'http://dassh.ac.uk/lifeforms/' (but is currently password protected for members of the Pelagics Working Group to test). The web front-end provides a user-friendly interface for users to access the lifeform extraction tool. Also available on the website is a versioned, current copy of the master list; the definitive, mapping of pelagic species to traits, with an associated Digital Object Identifier.



Screenshot of the DASSH lifeform tool.

Next Steps and Recommendations

To enable future expansion of the database, a template has been developed to allow users to submit new records which can also be integrated into the lifeform tool. An ingestion script for this template has also been developed which should allow the database to be automatically updated provided data is submitted using the developed form. Future work would involve testing and improving this script. The template is built around common, generic elements from the core biological MEDIN Data Guidelines to ensure the wideset possible interoperability. The template represents the common elements from the whole range of data providers. These map directly to data guideline elements should providers wish to submit fully compliant guideline-based datasets.

The resulting database and tools will be used for national, international, and EU-level policy and management assessment and reporting. In the next 5 years, at minimum, the data will support assessment of pelagic habitats for the upcoming OSPAR QSR and the next round of MSFD reporting. Further use of the data will require permission from the UK Pelagics Working Group and specific data holders. Standard DASSH permissions forms have been circulated to the data providers with an aim to make the raw data as open as possible.

Some datasets require further investigation for the use of specific lifeforms, such as gelatinous and small phytoplankton for the Continuous Plankton Recorder survey as they are not sampled effectively. These need to be excluded from the data tool, which can be detailed in the data agreement. Upon the completion of data agreements for each dataset provided, this web portal will be made available to wider users, where it can be utilised for pelagic monitoring and further scientific investigations.

DASSH – The Archive for Marine Species and Habitats Data

OSPAR region	*Ecohydrodynamic area*	*Institute - Region*	*Contacts*	*Lat*	*Lon*	*Sampling frequency*	*Parameters (Phytoplankton/ zooplankton)*	*Period*	*doi/usage/notes*
ALL	ALL?	MBA-Continuous Plankton Recorder UK/European Seas	David Johns			Monthly	P + Z	2004-2014	Access and re-use may be restricted, contact Data Owner. doi:10.7487/2016.263.1.1008
Region II: Greater North Sea	Indeterminate	MSS - Stonehaven	Elieen Bresnan Kathryn Cook	56.96	-2.13	Weekly	P + Z	1997-2014	Access and re-use may be restricted, contact Data Owner
	Region of freshwater influence	SEPA – Firth of Forth	Malcom Baptie (Elieen Bresnan)	56.02	-3.17	Monthly	P		
		Cefas - Warp, Dowsing, W Gabbard, Gabbard, Liv Bay 1, Liv Ba, Celtic Deep, Oyster Goup	Elisa Capuzzo, Veronique Creach, Michelle Devlin			3 monthly	P		Access and re-use may be restricted, contact Data Owner
	Seasonally stratified	PML – L4	Angus Atkinson Claire Widdicombe	50.25	-4.22	Weekly	P + Z	1992-2014 (P) 1988-2015 (Z)	Access and re-use may be restricted, contact Data Owner

	Seasonally stratified?	MSS – Scalloway (Shetlands)		60.18	-1.23	weekly	P	2002-present	
	Mixed inshore; tidal mixing	MSS – Scapa (Orkney Islands)		58.74	-3.04	weekly	P	2002-present	
	Permanently mixed	NLWKN (Lower Saxon Department for Water, Coastal and Nature Conservation) - island "Norderney" and sampled at high tide. Landesamt für Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein (LLUR)	Annika Grage	53.697033	7.165052	Weekly	P + Chla TEMP SUSP DOXY PH PSAL	1999-2014	Access and re-use may be restricted, contact Data Owner
	Indeterminate	SMHI (Swedish Meteorological and Hydrological Institute) - Swedish national monitoring data - Swedish West coast		56.6666	-12.1167	Monthly	P + Z	1986-2015 (P) 2007-2015 (Z)	Access and re-use may be restricted, contact Data Owner
	Seasonally stratified	AFBI – western Irish Sea	Cordula Scherer, Matt Service	53.78	-5.64	Monthly	P	2008-2010	Access and re-use may be restricted, contact Data Owner
	Permanently mixed	AFBI LBy06 – proposed new site				3 monthly			

	Predominantly haline stratification	SEPA- Inner Firth of Clyde		55.94	-4.89	Monthly	P		
Region III: Celtic Seas	Complex seasonality	MSS – Loch Ewe	Elieen Bresnan Kathryn Cook	57.84	-5.61	Weekly	P + Z	2002-2014	Access and re-use may be restricted, contact Data Owner
	Predominantly haline stratification	SAMS – LY1 (Firth of Lorne/Loch Linnhe)	Paul Tett	56.48	-5.5	Monthly	P	2000-present	
	Region of Freshwater Influence	EA – ECMAS - Inner Bristol Channel Minehead	Mike Best			Monthly	P	2010-2014	Access and re-use may be restricted, contact Data Owner
Region IV: Bay of Biscay and Iberian Coast?	Indeterminate	INSTITUTO ESPAÑOL DE OCEANOGRAFÍA - Radiales				Monthly	P + Z		Access and re-use may be restricted, contact Data Owner
Region V: Wider Atlantic									
		Department of Bioscience Aarhus University - Danish	Hans Jakobsen & Eva Friis Møller						Access and re-use may be restricted, contact Data Owner. Not recognized data format - ICES format?

		Water Framework Directive (WFD) - Phyto over 430 sites	Mike Best	430 sites	430 sites	Not always consistent	P	2007-2015	Access and re-use may be restricted, contact Data Owner

Datasets included in the ingestion in green. Other datasets require transformation prior to inclusion and will be integrated as funding allows.