SeaDataNet

PAN-EUROPEAN INFRASTRUCTURE FOR OCEAN & MARINE DATA MANAGEMENT

FORMAT DOCUMENTATION

Recommended data transport format for biological data in the framework of SeaDataNet

Deliverable D8.4b

Project Acronym : SeaDataNet II
Project Full Title : SeaDataNet II: Pan-European infrastructure for ocean and marine data management
Grant Agreement Number : 283607

SEVENTH FRAMEWORK PROGRAMME

Deliverable number	Short Title
D8.4 (part B)	Transport format for biological data in SeaDataNet
Long title	
Recommended data transport format for biological data in the framework of SeaDataNet	
Short description	
This document describes the proposed data transport format for biological data in the framework of SeaDataNet.	
Author	**Working group**
K. Deneudt	WP8

History

Version	Authors	Date	Comments
1.0	K. Deneudt	28/11/2013	Creation

Content

1. Introduction

One of the objectives of SeaDataNet II is to undertake actions to make SeaDataNet better fit for handling marine biological data sets and establishing interoperability with biology infrastructure developments. Based on an analysis of present biology data standards and initiatives, such as the OBIS, GBIF, TDWG and WoRMS standards, a recommended format for data transport of biological data is developed. This document describes the proposed format.

2. Objectives

The format should enable National Oceanographic Data Centers (NODC's) to make biological data accessible using SeaDataNet infrastructure and should make it possible for NODC's to use SeaDataNet to exchange biological data and to contribute to initiatives like (Eur)OBIS and GBIF.

This general purpose results in two specific requirements for the format:

- The format should be a general and higher level format without necessarily containing all specifics of each data type, but rather focusing on common information elements for marine biological data.
- At the same time the format needs to be sufficiently flexible/extendable to be applicable for at least part of the variety of biological data NODC's are managing.
- It should be possible to derive OBIS or Darwin Core compatible datasets from the format.

The format should be self-describing, in the sense that all information needed to interpret the data should be included in the file format or available through links to vocabularies or term lists that are part of the format.

While the primary target is data managers, it would be preferable to have a simple flat and human readable format that can as easily be created by scientists themselves starting from their original biological datasets which often reside in spreadsheets.

3. Format documentation for biology data

Accommodating data in the SeaDataNet infrastructure requires a detailed description of the data in the the SDN metadata formats (CDI, CSR, EDMED, EDMERP, EDIOS) and storage of the data in one of the data transport formats (ODV, MedAtlas, NetCDF files).

For providing data access to biology data through the SDN portal the creation of CDI XML files storing the metadata and corresponding ODV files is the chosen approach. The CDI XML files for biology metadata are no different from the general CDI files. For the data itself, a specific variant of the ODV files for biology data is available and described below.

3.1. CDI metadata

Since CDI XML files for biology metadata are not really different from other CDI files we can refer for the creation of these files to existing documentation on the CDI (ISO193139 model). http://www.seadatanet.org/Standards-Software/Metadata-formats/CDI

For certain types of biology data specific optional CDI fields might be required. For example:

TRACKS (Curves)	When describing line tracks by eg. demersal fishing, this optional field can be used to describe the fishing track.
AREAS (Surfaces)	In case areas are surveyed instead of point stations, this optional field can be used to describe the multi-surface or polygon of each area entity.

The use of an EDMED REFERENCE in the CDI is highly recommended for biology data. EDMED is a dataset catalogue that can be used to create discovery metadata at a higher aggregate level. This field can be used to refer to the appropriate EDMED code referring to an item in the EDMED catalogue where a higher level description of the data collection can be found. Instructions for creating an entry for the EDMED catalogue can be found at http://www.seadatanet.org/Standards-Software/Metadata-formats/EDMED.

3.2. ODV biology variant

3.2.1. Mandatory fields

The general format of the SeaDataNet ODV import format is described at http://www.seadatanet.org/Standards-Software/Data-Transport-Formats.

The biology variant of the ODV file deviates from the general format in a number of ways. The format uses specific mandatory data parameter columns in addition to the mandatory elements already described in the Data Transport Format Manual. These biology mandatory fields are described below. Each of them have corresponding P01 entries that should be included in the header.

MinimumDepthOfObservation/ MaximumDepthOfObservation	Describes the minimum and maximum observation depths. Refers to: www.seadatanet.org/urnurl/SDN:P01::MINWDIST Refers to: www.seadatanet.org/urnurl/SDN:P01::MAXWDIST

SampleID	A unique identifier for a unit of material, or other medium of observation, gathered at a discrete point in time, that serves as a sample for a survey of biological occurrence (presence, quantification, absence, or derived value). The definition of a sample and assignment of IDs is controlled by the data originator. If applicable, details about how sample are defined and IDs assigned will be provided in metadata [original term from IOOSBiology]. *Refers to:* www.seadatanet.org/urnurl/SDN:P01::SAMPID01
SamplingEffort [unit]	The amount of effort expended during an Event. (It can be a volume (e.g. for a phytoplankton sample), a linear distance (e.g. for a visual transect or net haul), a surface area (e.g. for a benthic core), etc) [original term from DwC]. *Refers to:* www.seadatanet.org/urnurl/SDN:P01::LENTRACK (for length) *Refers to:* www.seadatanet.org/urnurl/SDN:P01::VOLWBSMP (for volume) *Refers to:* www.seadatanet.org/urnurl/SDN:P01::AREABEDS (for area)
SubsampleID	If a sample is further divided by the data originator for purposes of organization, handling, or analysis, and if the data originator wants to maintain the identity of the subsample, capture the ID of the subsample here. Explain in metadata if applicable [original term from IOOSBiology]. *Refers to:* www.seadatanet.org/urnurl/SDN:P01::SSAMID01
SubSamplingCoefficient	A decimal coefficient lower or equal to 1 that indicates the proportion of the Sample that is represented in the Subsample. This value allows to rescale the counted (and measured) individuals to the total sampling effort applied at Sample level. If no proportional subsampling occurred this value should equal 1. *Refers to:* www.seadatanet.org/urnurl/SDN:P01::SSAMC01
ScientificName	The full name of lowest level taxon the Cataloged Item can be identified as a member of; includes genus, specific epithet, and subspecific epithet (zool.) or infraspecific rank abbreviation, and infraspecific epithet (bot.) Use name of suprageneric taxon (e.g., family name) if Cataloged Item cannot be identified to genus, species, or infraspecific taxon. Check spelling of names against World Register of Marine Species; see notes with the schema for further details on taxonomy [original term from OBIS]. Refers to: www.seadatanet.org/urnurl/SDN:P01::SCNAME01
ScientificNameID	An identifier for the nomenclatural (not taxonomic) details of a scientific name [original term from OBIS]. Refers to: www.seadatanet.org/urnurl/SDN:P01::SNANID01
Sex	The sex of a specimen or collected/observed individual(s). The domain should be a controlled set of terms (codes) based on community consensus. Proposed values: M=Male; F=Female; H=Hermaphrodite; I=Indeterminate (examined but could not be

	determined; U=Unkown (not examined); T=Transitional (between sexes; useful for sequential hermaphrodites); B = Both Male and Female [original term from OBIS]. Refers to: www.seadatanet.org/urnurl/SDN:P01::ENTSEX01
LifeStage	Indicates the life stage present. Will require developing a controlled vocabulary. Can include multiple stages for a lot with multiple individuals [original term from OBIS]. Refers to: www.seadatanet.org/urnurl/SDN:P01::LSTAGE01
ObservedIndividualCount	For the taxon under consideration, give the number of individuals that was found in the (sub)sample described by (sub)SampleID. The ObservedIndividualCount represents the number of individuals in the (sub)sample uniquely identified by the (sub)SampleID. [original term from OBIS]. Refers to: www.seadatanet.org/urnurl/SDN:P01::OCOUNT01

As is the case for all data value columns in the ODV format, all the above data fields are paired with qualifying flags using the SeaDataNet vocabulary for qualifying flags (QV:SEADATANET). The Qualifying flag codes are available at http://vocab.nerc.ac.uk/collection/L20/current.

3.2.2. Optional fields

Like the general ODV format the biology variant allows the use of additional non-mandatory data parameter columns chosen from available terms in the SeaDataNet P01 and P06, and relevant for specific data types.

Examples of these additional fields are:

- DensityPerUnitEffort [1/m²]
www.seadatanet.org/urnurl/SDN:P01::SDBIOL02
(definition: Abundance of unspecified biological entity per unit area of the bed)
www.seadatanet.org/urnurl/SDN:P06:PMSQ
(definition: per square metre)

- Lead_per_unit_wet_weight_of_biota
http://www.seadatanet.org/urnurl/SDN:P01::PBBIOTUK
(definition: Concentration of lead {Pb} per unit wet weight of biota)
http://www.seadatanet.org/urnurl/SDN:P06::UMKG
(definition: Milligrams per kilogram)

- ObservedIndividualLength
http://www.seadatanet.org/urnurl/SDN:P01::OBSINDLX
(definition : Length of unspecified biological entity)
http://www.seadatanet.org/urnurl/SDN:P06::ULCM
(definition : Centimetres)

3.2.3. Semantic header

The semantic header of the ODV biology variant follows the general ODV semantic header rules as outlined in the Data Transport Formats manual at http://www.seadatanet.org/Standards-Software/Data-Transport-Formats.

The semantic header is needed as a mapping that provides a reference for every data column heading label to standardized vocabulary concepts.

Notice that in the newest version an <instrument> element is allowed in addition to the <object>, <subject>, <units> element. The use of this <instrument> element referring to the L221 vocabulary is highly recommended for biology data. http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx/

4. Examples of biology data

A number of example files have been drawn up demonstrating how to deal with a diverse range of biological datasets. The examples show how biology data can be dealt with using:

CDI

- Mandatory fields CDI (ISO19139 model)
- Specific optional fields CDI (ISO19139 model)

ODV biology variant

- Mandatory fields
- Optional fields

For purposes of demonstration and fitness for use assessment the examples are worked out in spreadsheet (BioODV_1.0). This allows to add comments and present a more human readable version of the format. In reality analogous CDI XML files and ODV txt files should be produced.

- **Example 1**: Grab/core benthos community data (BioODV_1.0_MacroB) - This example demonstrates how to cover both density and biomass of infaunal (in this case macrobenthos) community data.

CDI mandatory fields	CDI optional fields	ODVbio mandatory fields	ODVbio optional fields
All mandatory CDI fields	No specific optional CDI fields	Cruise	DensityPerUnitEffort [unit]*
		Station	AFDWBiomassPerUnitEffort [unit]*
		Type	AFDWBiomass [unit]*
		yyyy-mm-ddThh:mm:ss.sss	…
		Longitude [degrees_east]	
		Latitude [degrees_north]	
		LOCAL_CDI_ID	
		EDMO_code	
		Bot. Depth [m]	
		MinimumObservationDepth	
		MaximumObservationDepth	
		SampleID*	
		SamplingEffort [unit]*	
		SubSampleID*	
		SubSamplingCoefficient*	
		ScientificName*	
		ScientificNameID*	
		Sex*	
		LifeStage*	
		ObservedIndividualCount*	

(*) fields indicated require paired QC flag column

- **Example 2**: Vertical Profile zooplankton community data (BioODV_1.0_ZooP) - This example demonstrates how to cover zooplankton data from vertical net sampling of different depth zones.

CDI mandatory fields	CDI optional fields	ODVbio mandatory fields	ODVbio optional fields
All mandatory CDI fields	No specific optional CDI fields	Cruise	DensityPerUnitEffort [*unit*]*
		Station	…
		Type	
		yyyy-mm-ddThh:mm:ss.sss	
		Longitude [degrees_east]	
		Latitude [degrees_north]	
		LOCAL_CDI_ID	
		EDMO_code	
		Bot. Depth [m]	
		MinimumObservationDepth	
		MaximumObservationDepth	
		SampleID*	
		SamplingEffort [*unit*]*	
		SubSampleID*	
		SubSamplingCoefficient*	
		ScientificName*	
		ScientificNameID*	
		Sex*	
		LifeStage*	
		ObservedIndividualCount*	

(*) fields indicated require paired QC flaq column

- **Example 3**: Fish trawl data (BioODV_1.0_DemFish) - This specific example demonstrates how to cover demersal fish population data resulting from towed net fish trawls. The example demonstrates how to deal with subsampling storing information on the applied SubSamplingCoefficient and storing identifiers for both sample and subsample. The example includes both real counts of fish and derived densities per unit effort. Furthermore exemplary cases for including length-frequency data (counts and densities for different size classes), individual fish length measurements, and age information are available.

CDI mandatory fields	CDI optional fields	ODVbio mandatory fields	ODVbio optional fields
All mandatory CDI fields	Tracks (Curves)	Cruise	SubSamplingCoefficient*
		Station	ObservedMinLength [*unit*]*
		Type	ObservedMaxLength [*unit*]*
		yyyy-mm-ddThh:mm:ss.sss	ObservedIndividualLength [*unit*]*
		Longitude [degrees_east]	DensityPerUnitEffort [*unit*]*
		Latitude [degrees_north]	Age [*unit]**
		LOCAL_CDI_ID	SampleDuration [*unit*]*
		EDMO_code	
		Bot. Depth [m]	
		MinimumObservationDepth	
		MaximumObservationDepth	
		SampleID*	
		SamplingEffort [*unit*]*	
		SubSampleID*	
		SubSamplingCoefficient*	
		ScientificName*	
		ScientificNameID*	
		Sex*	
		LifeStage*	
		ObservedIndividualCount*	

(*) fields indicated require paired QC flaq column

- **Example 4**: Biota pollutant concentration data (BioODV_1.0_BiotaPol) - This example demonstrates how to cover data on pollutant concentrations in biota.

CDI mandatory fields	CDI optional fields	ODVbio mandatory fields	ODVbio optional fields
All mandatory CDI fields	*No specific optional CDI fields*	Cruise	Lead_per_unit_wet_weight_of_biota[*unit*]*
		Station	…
		Type	
		yyyy-mm-ddThh:mm:ss.sss	
		Longitude [degrees_east]	
		Latitude [degrees_north]	
		LOCAL_CDI_ID	
		EDMO_code	
		Bot. Depth [m]	
		MinimumObservationDepth	
		MaximumObservationDepth	
		SampleID*	
		SamplingEffort [*unit*]*	
		SubSampleID*	
		SubSamplingCoefficient*	
		ScientificName*	
		ScientificNameID*	
		Sex*	
		LifeStage*	
		ObservedIndividualCount*	

(*) fields indicated require paired QC flaq column