Data Management Plan of the Baltic Data Flows Project

May 2021

Contents

1. Introduction

Environmental monitoring data collected and processed under activities of the Baltic Data Flows (BDF) project are part of the existing HELCOM monitoring data collection system agreed by Contracting Parties of Helsinki Convention and adhere to the HELCOM Monitoring and Assessment Strategy. Thus, access to data and data use policy described there will apply to the environmental monitoring data processed in this Action. For data, software and documents generated and used in the Action, this Data Management Plan (DMP) has been developed to ensure data availability and security within the consortium (for the Action duration and beyond).

The purpose of this DMP is to serve as a guidance document for BDF partners for data collection, access, storage, protection, preservation and maintenance, metadata generation, retention, and destruction of data, and it ensures that BDF partners will comply with national and EU legislations throughout the duration of the BDF project. In addition, for externally derived data products, e.g., earth observation data, data use conditions set by the data providers will be followed.

The DMP aims at summarizing guidelines, agreements, policies, and best practices for data management for the whole data lifecycle (figure 1) as well as at providing an effective framework to ensure comprehensive collecting and handling of the data used in the project. The DMP will be consistent with FAIR (findable, accessible, interoperable, and reusable) data principles.
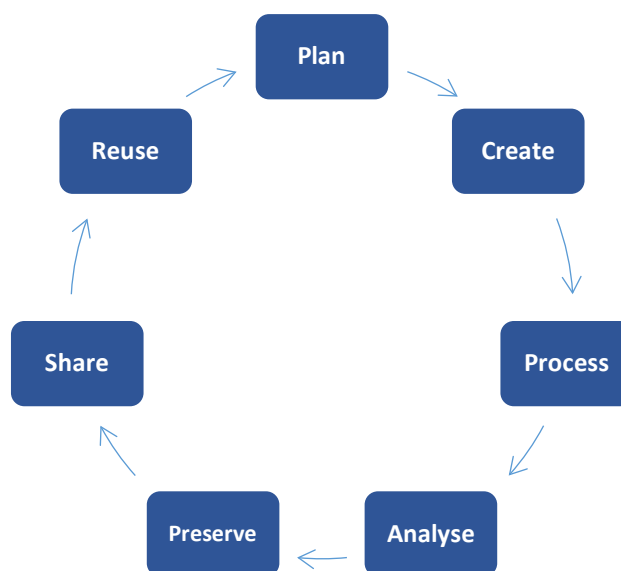

Figure 1: Data lifecycle

2. Project summary

The Baltic Data Flows project, co-financed by the Connecting Europe Facility of the European Union, seeks to enhance the sharing and harmonisation of data on the marine environment originating from existing sea monitoring programmes, and to move towards service-based data sharing.

In particular, open datasets will be made available by HELCOM to a wider community, such as European open data ecosystem, researchers, NGOs and private sector, in order to benefit from the availability of harmonised environmental data. Wider dissemination is achieved by sharing HELCOM metadata records to European Data Portal.

Baltic Data Flows will improve the capacity building of the national environmental data hosting organisations and providers of the consortium, in terms of quality control and solutions to make harmonised environmental data available. Members of the consortium will build and enhance their ICT infrastructure to support better the data sharing process. Furthermore, data harvesting systems based on Application Programming Interfaces (APIs) will be developed with the aim to automatically integrate national datasets into a combined and harmonised regional dataset for partner institutes SMHI and SYKE for the environmental monitoring data that is reported to ICES.

Furthermore, Baltic Data Flows will further develop tools and indicator data flows for eutrophication, hazardous substances, and biodiversity assessment. This will be carried out by further developing existing data flows, data view and aggregation tools at data host ICES and by further defining the indicator data needs. The focus of the work is indicators related to eutrophication, hazardous substances and biological community data reported to ICES (Seasonal succession of dominating phytoplankton groups, Zooplankton mean size and total stock (MSTS), State of the soft-bottom macrofauna community).

The work will include also further development of assessment tools and tool outputs in close cooperation with relevant expert groups / indicator leads and ultimately State & Conservation. For the biodiversity related parts, synergy with HELCOM BLUES project outputs on Biodiversity assessment Tool (BEAT) will be explored.

The project will run from October 2020 to October 2023. Project partners include HELCOM, ICES, LHEI, SMHI, Spatineo inc., Stockholm University, and SYKE.

3. Data collection and generation

Data collection refers to the process of collecting data from sources outside of the project, whilst data generation refers to generating new data through project activities. The BDF project facilitates data collection from a range of external sources and generates new data by analyzing and using the collected data for assessing the status of marine environment for environmental policy instruments such as HELCOM and for EU Marine Strategy Directive. One of the principle aims of the project is to promote open datasets to ensure public data is widely used and available for diverse audience types. This will encompass a variety of datasets with particular focus on marine eutrophication, hazardous substances and biodiversity data, through the objectives of activities 4, 5, and 6 outlined in this section.

Partners of the project are data holders and national competent authorities collecting and hosting marine data and sharing the data to international and EU processes.

Data from external sources will be collected both from primary and secondary sources. Primary data is collected/made available directly by project members. Secondary data is existing data collected by a party other than the project, such as registry data or survey data. The project will utilize existing and archived environmental data, models, scripts, results from other projects and studies as well as satellite services, measurements, and observations.

With respect to the different data types that the project will collect and generate, a distinction will be made between personal data, sensitive data, and non-personal data. According to the European Commission, personal data is *"any information relating to an identified or identifiable living individual."* Different pieces of information collected, which can lead to the identification of a particular person, also constitutes personal data. In accordance with GDPR Article 9, sensitive data refers to *"special categories of personal data"*, defined as *"data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation"*. Non-personal data refers to data that does not constitute personal data under Article 4 of GDPR.

The BDF project will collect and generate environmental data relating to the Baltic marine environment. The nature of this data is non-personal. However, metadata records related to these datasets may contain personal and sensitive data, so it is therefore important to consider EU regulations within this project.

All data in the BDF project will be in digital format. The BDF DMP complies with the EU's guidelines regarding the DMP. Therefore, this document and, more specifically the following sections, will address and describe for each data set collected, processed and/or generated in the project the following characteristics: dataset description, reference and name, standards and metadata, data sharing, archiving and preservation.

3.2. Eutrophication data

HELCOM eutrophication core indicators evaluate the status of the Baltic Sea as reflected by concentrations of nutrients in the water, water clarity and algae blooms as well as oxygen concentration. The core indicators are commonly agreed tools to follow up on the progress made towards the Baltic Sea Action Plan overall goal of achieving good environmental status in the Baltic Sea by 2021.

Activity 4 of the project will generate data products on the marine environment by combining raw data from different monitoring methods (classical in-situ monitoring, ferrybox - automatic measurement devices on board commercial ships, and earth observation data). The activity analyses and makes use of suitable data products and services available from EU Copernicus programme Sentinel Satellite series (e.g. SYKE downstream service for the Baltic Sea, Copernicus Marine Environment Monitoring Service – CMEMS). This activity will concern data on eutrophication indicators such as chlorophyll-a, cyanobacterial bloom index indicators. The activity is led by SYKE with the participation of SMHI, ICES and HELCOM.

A data flow model highlighting the planned data flows for receiving ferry box data between partners has been generated as shown in figure 2. Figure 2 shows the primary sources of eutrophication data in the grey boxes. The 'AlgaBase', an internal database provided by SYKE, collects flow-through from ferry boxes. SMHI are sharing flow-through ferry box data via 'CMEMS' and bottle samples via 'SHARKWEB'. There are also other ferry box data providers such as commercial cruise liners. This data is then processed to develop the indicator calculation that is used for assessment (HEAT model).

Consideration will be applied for proper use, citation and reference to original data sources on original data from CMEMS and national data providers.
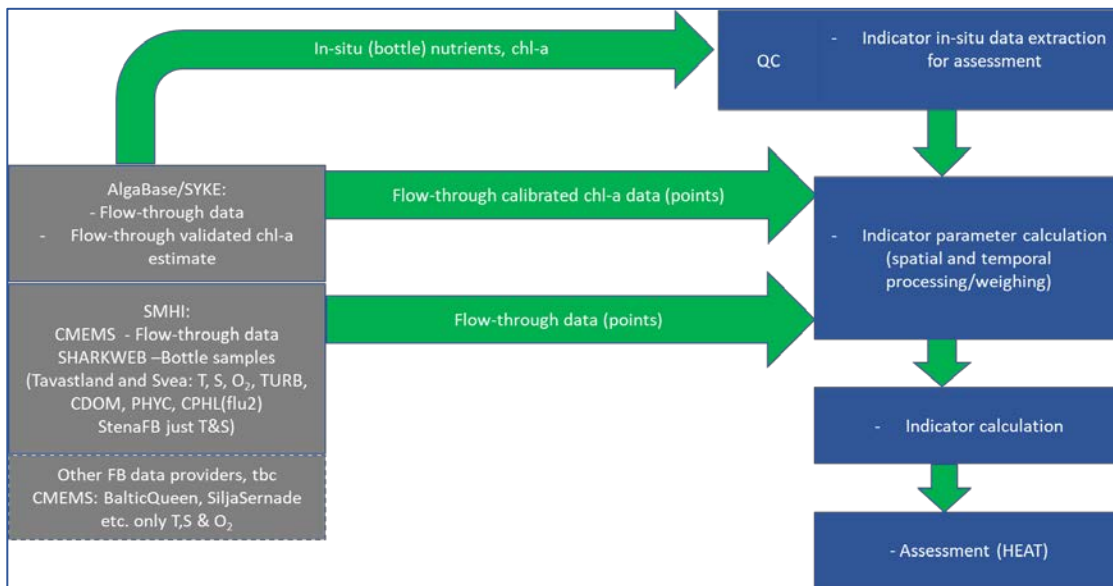
Baltic
DataFlows

Figure 3: Ferry box data flow model

A similar data flow model has been generated for earth observation data, as shown in figure 3. Figure 3 shows the primary sources of earth observation data in the grey boxes. The Tarkka database developed by SYKE is a resource for open satellite data, based on a 20x20km grid from Sentinel 2 satellite data. This will provide the chl-a estimates, and the Sentinel 3 and MERIS will provide the cyanobacteria data. There are also other earth observation data sources that will be defined later in this project. This data is then processed to develop the indicator calculation that is used for assessment (HEAT model). When the new datatypes are fully integrated to the data flow models the next step will be to generate eutrophication indicator data products by combining the three data types (in-situ, ferry box and earth observation). The task will be performed with close cooperation and dialogue with the relevant HELCOM expert group (HELCOM IN-Eutrophication and State & Conservation).
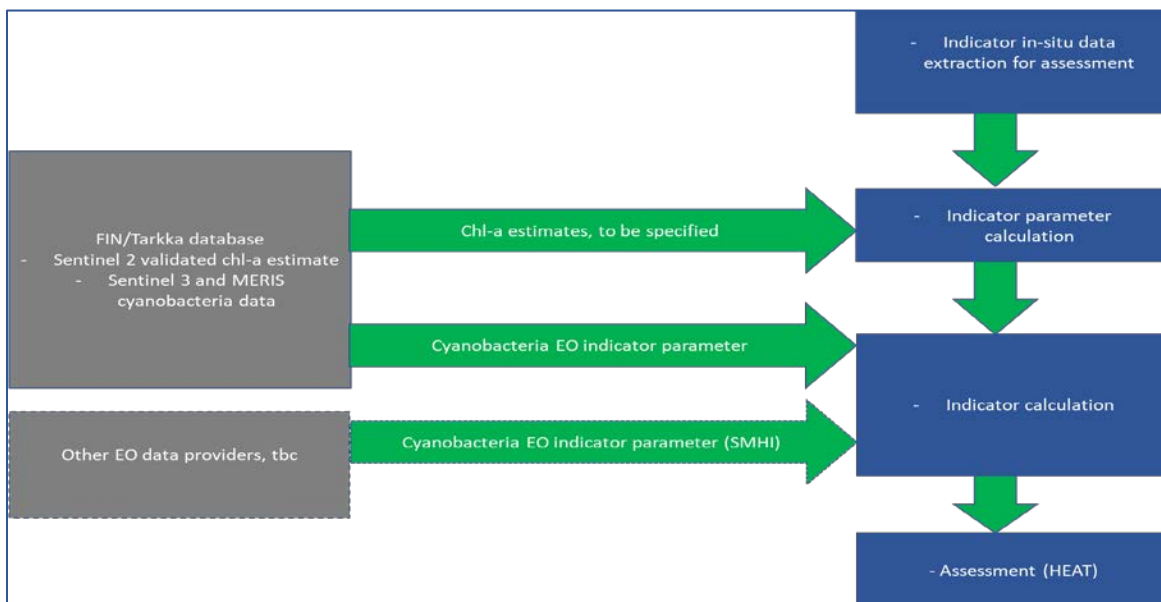

Figure 2: Earth observation data flow model

3.3. Hazardous substances data

HELCOM hazardous substance core indicators evaluate the status of the Baltic Sea as reflected by concentrations of hazardous substances in animals, sediment, and seawater as well as effects displayed by animals. The core indicators are commonly agreed tools to follow up on the progress made towards the Baltic Sea Action Plan overall goal of achieving good environmental status in the Baltic Sea by 2021. Activity 5 of the project enhances creation of harmonised content and data products related to hazardous substances in the marine environment. This activity will be led by HELCOM with the participation of SYKE, LHEI and ICES. This activity is composed of 3 separate tasks that collect and generate data on hazardous substances.

Task 5.1. involves reviewing and developing a methodology for hazardous substances indicator calculation. This task includes the review of used supporting parameters (metadata on samples), conversion factors, and assessment parameters within current indicators; the comparison of reported monitoring data vs. data requirements for indicators in order to make sure required metadata on samples (e.g. matrices, dry weights, etc.) is collected during monitoring and laboratory analysis steps, and that it is stored to the national database and further reported/shared to the regional database; test runs of the above-mentioned subtasks to compare the impact on the overall assessment and data availability at the indicator evaluation and data visualisation stage;

Task 5.2. will further develop the existing assessment methodology and tools (CHASE). It includes outlining the development needs of the CHASE assessment tool, planning additional HELCOM indicators and the confidence assessment to be part of the assessment workflow, and improved incorporation of Radioactive Substances evaluation and adjustment of the assessment of the confidence to address independent substances (e.g., metals). Figure 4 displays the planned assessment workflow on completion of Activity 5 showing acceptance levels 1-3 with the purple arrows from reporting into the CHASE.
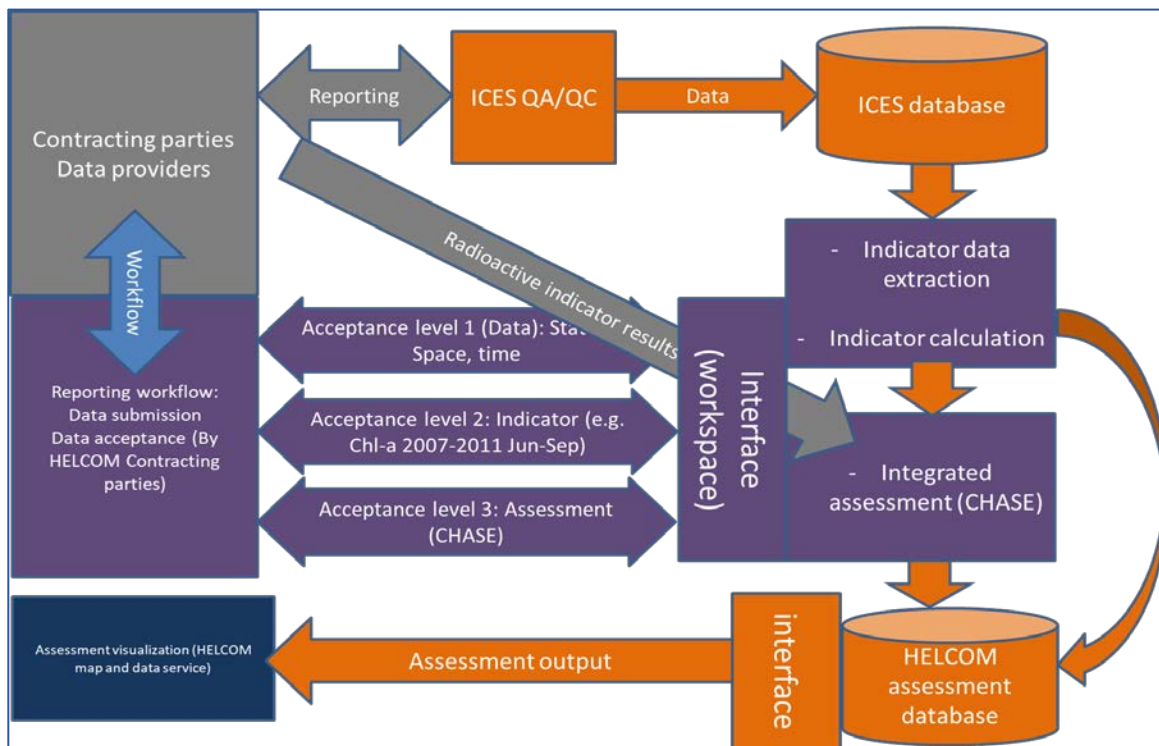

Figure 4: Hazardous substances workflow

Task 5.3 will further develop and implement the existing hazardous substances online assessment tool hosted by ICES according to the defined assessment methodology and required outputs / data products of task 5.2. It includes accessibility and presentation of all data and evaluation outputs via an online assessment tool; the incorporation and visualisation of a Radioactive Substances assessment indicator and the inclusion of new substances and their assessments (e.g., copper in the marine environment); and improved confidence evaluation of the indicators based on data availability, temporal, and spatial coverage (per assessment area).

3.4. Zooplankton, Phytoplankton and Zoobenthos data

HELCOM biodiversity core indicators evaluate the status of the Baltic Sea as reflected by marine mammals, seabirds, fish, benthic biotopes and pelagic plankton communities. The core indicators are commonly agreed tools to follow up on the progress made towards the Baltic Sea Action Plan overall goal of achieving good environmental status in the Baltic Sea by 2021.

Activity 6 will develop data harmonisation methods to facilitate the creation of harmonised cross-border indicator data products to support the assessment of the status of biodiversity in the marine environment using biological community data. This features zooplankton, phytoplankton and zoobenthos data. Overall, activity 6 will be led by SYKE with the participation of SMHI, LHEI, SU, ICES and HELCOM. This activity is composed of 3 separate tasks:

- 6.1. Reviewing and updating indicator data requirements for harmonized indicator data products
- 6.2. Reviewing and developing methodology for indicator calculation.
- 6.3. Developing methodology for integrated biodiversity assessment and assessment tool

Figure 5 presents the phytoplankton indicator workflow, and figure 6 presents the zooplankton indicator workflow, each showing how data is combined from a variety of sources and feed into the BEAT assessment tool.
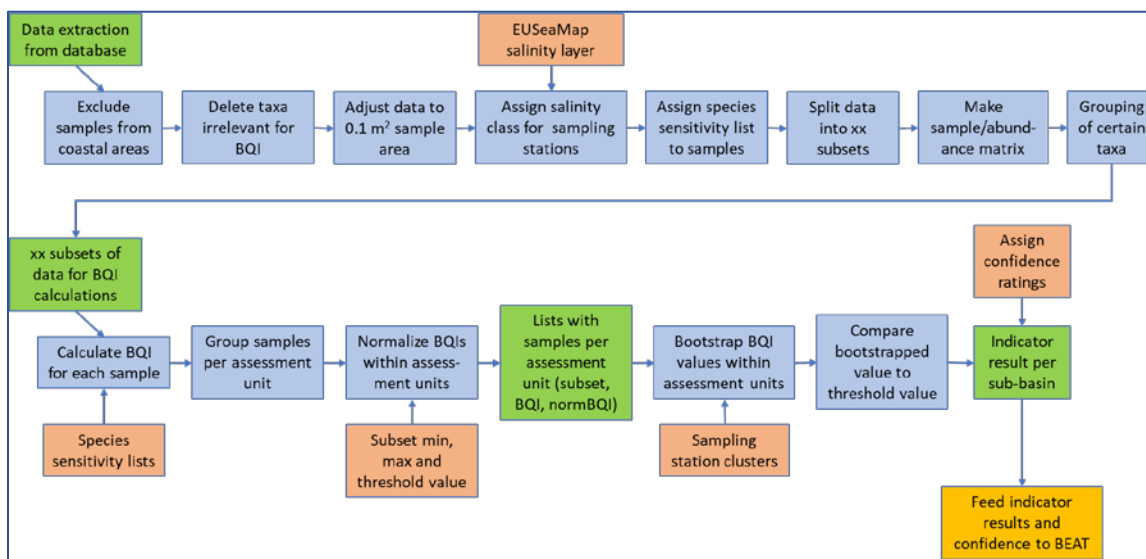

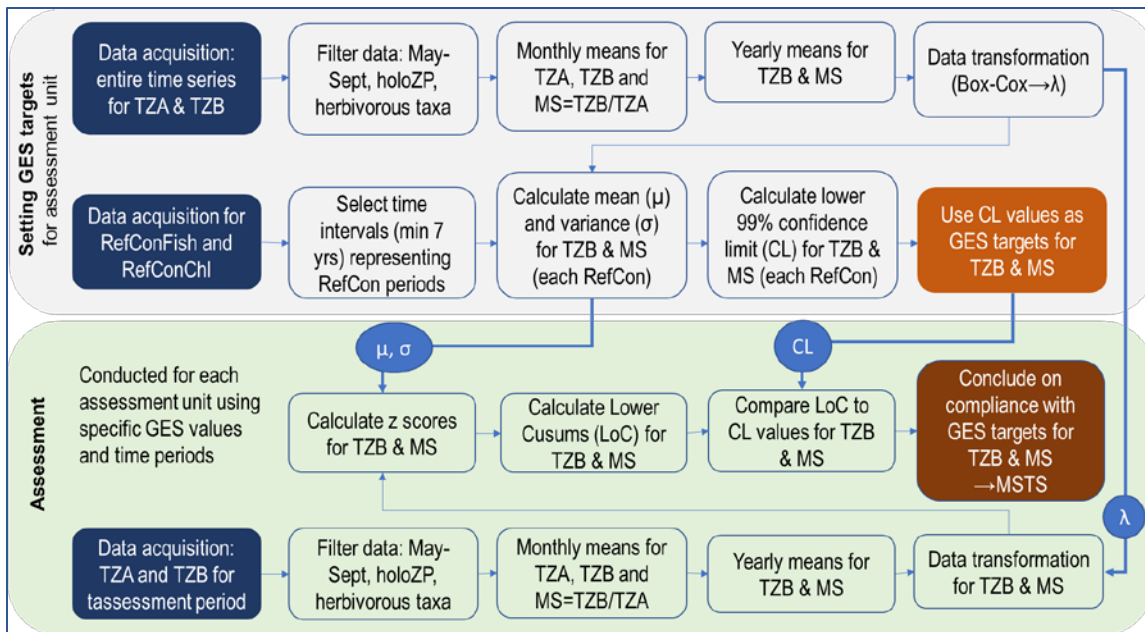Figure 5: Phytoplankton indicator workflow

Figure 6: Zooplankton indicator workflow

4. FAIR Data Management

This project will uphold the FAIR Principles throughout all activities that collect and generate data as listed in section 3 of this DMP. For each activity listed in section 3 a separate sub-task dedicated to 'making assessment data products FAIR' is integral to the GA. In order to fully understand and apply the FAIR Principles this DMP has referenced the 'Final Report and Action Plan from the European Commission Expert Group on FAIR Data'

4.1.   Making data findable

Data is Findable when described by rich metadata and registered or indexed in a searchable resource that is known and accessible to potential users. Additionally, a unique and persistent identifier should be assigned such that the data can be unequivocally referenced and cited in communications. The identifier enables persistent linkages to be established between the data, metadata and other related materials in order to assist data discovery and reuse.

The BDF project will achieve this for all data collected and generated for data types identified in section 3. For assessment data products under activities 4, 5 and 6, a procedure for assigning a digital object identifier (DOI) to the assessment of datasets will be established. In addition, metadata records for harmonized regional datasets will be created by making the dataset accessible and by assigning a DOI to the dataset. This ensures that consecutive work done using these datasets can be conducted by downloading them from data repositories without asking individual permissions from authors and they can be referenced with DOI links.

The BDF project Website will provide main outcomes and updates. Data discoverability will be achieved by search keywords that describe BDF's data and outcomes and are related to standards from publishers or repositories.

Data products published in HELCOM Metadata Catalog can be referenced and accessed from defined URL enabling referencing this data e.g. in data, presentation or in other usage. The metadata records can be also harvested to

other metadata catalogues or information sources using currently available CSW interface and also via DCAT-AP interface fostering the accessibility and reuse of the data.

Metadata template applied in the catalogue contains fields described in [ISO 19115 metadata standard](#) and EU INSPIRE metadata Implementing Rules (e.g. abstract, lineage etc.). GEMET (GEneral Multilingual Environmental Thesaurus) is used for keywords for categorizing datasets in [HELCOM Metadata Catalog](#).

4.2. Making data accessible

Accessible data objects can be obtained by humans and machines upon authorisation and through a defined and universally implementable protocol. Project data dissemination will feature a variety of outputs, for example, selected sets of output spatial data products (maps) are accessible as a service via standard spatial data service interfaces, e.g. OGC WMS for viewing and CSW for metadata descriptions. Output datasets are downloadable from each metadata record in a standard and widely used spatial data output format (e.g. NetCDF, TIFF, shapefile).

More specifically under Activity 4 metadata records will be created for the relevant in-situ, ferrybox and earth observation datasets in national open data portal by SYKE and SMHI, utilizing existing national platforms. In addition, under Activities 5 and 6 metadata records of national datasets used for assessment will be published and hosted by Action partners in national open data portals and in the [European Data Portal](#). In addition, all scripts used for indicator calculation and integrated assessment will be published on [HELCOM GitHub](#) account under an open public license, where applicable. This also addresses the following FAIR principles of Interoperability and Reuse.

4.3. Making data interoperable

Interoperable data and metadata are described in the FAIR principles as those that use a formal, accessible, shared, and broadly applicable language for knowledge representation. They use vocabularies which themselves follow the FAIR principles, and they include qualified references to other data or metadata. Strict standardization of the project output is the key for interoperability. In general, data interoperability is a priority for data shared among BDF partners as well as with third parties either through agreements or open access.

BDF data will be created, managed, and stored appropriately and in line with applicable legislation. To enhance interoperability of BDF data, data standards will be used for the organization, name, code and format of BDF data. In case of data deposited in repositories, specific repository standards will be used.

Metadata constitutes important information or description of the project data, as it contains information on the purpose, origin, time references, geographic locations, creator, access conditions and terms of use. Metadata facilitates finding and working with particular instances of data. Metadata should consist of information related to parameters used, temporal and spatial characteristics, methodology, objective(s) of the work, quality of the data as well as contact information of organization focal point responsible for the project. Metadata will also provide information of any kind of data handling related to data manipulation or any change occurred.

4.4. Re-use

For data to be Reusable, the FAIR principles assert the need for rich metadata and documentation that meet relevant community standards and provide information about provenance. This covers reporting how data was created and information about data reduction or transformation processes to make data more usable and understandable.

Baltic DataFlows

BDF data, including metadata, either deposited in repositories or shared through (GDPR and the EU) open access policy will be disseminated to improve reusability (such as repeat experiments, measurements, methodologies, assessments, etc.) for a reasonable period. Data reusability for an indefinite period will be ensured by having open file formats for the final shared datasets rather than proprietary formats. For some repositories, two key requirements, which allow for the re-usability of the data are: a) users' registration and b) acceptance of the terms and conditions such as proper citation and reference of the data source.

Data products resulting from the project will be made publicly available via HELCOM Map and Data Service and the HELCOM Metadata Catalog, where data use conditions for each dataset is described on metadata record level. To provide free access to data on the environment, the service is open access, meaning there is no need for user registration. Users entering the services are required to acknowledge the dataset specific data use conditions when entering the service. The metadata records can be harvested to other metadata catalogues e.g., thematic, national, or international open data portals fostering the accessibility and reuse of the data.

The reuse of data will be further supported by establishing a procedure for assigning a digital object identifier (DOI) to the assessment of datasets. The project will make available harmonised regional assessment products by creating metadata records for the dataset, by making the dataset accessible for assessment and by assigning a DOI to the dataset. This is related to activity 2, task 2.1, Implementation of DCAT-AP/INSPIRE in regional metadata catalogues and harvesting by the EDP. In this task, regional metadata catalogues (HELCOM and ICES) will be upgraded to be DCAT-AP compliant and to the newest version of Geonetwork, as needed. Under this task, also machine readability of data use conditions of metadata records will be developed to be compatible with the EDP guidance and requirements.

5. Data security

5.1 General

All data collected and generated (i.e. datasets, metadata, etc.) throughout the project will be stored in HELCOM and project partners' institutional servers with a reliable back-up system. HELCOM and project partners are responsible for ensuring that all data are stored securely and safely and in full compliance with the data protection regulations of European Union (i.e. GDPR). Moreover, all partners are obliged to keep raw data (1st category data) in their original form, and they will only analyse copies of them. Raw data shall be within partners' responsibility. Every partner must save all kinds of scripts and/or codes of data processing, so that all steps of data processing can be traced easily.

All partners are obliged to handle and process data, especially modelling data, according to local data handling procedures in their institution (2nd category data). This category of processed data (i.e. processed subsets of raw data) will be disseminated within BDF Consortium and they should be available for the lifetime of the project.

Finally, the final processed data (3rd category data) which are the data obtained from the further analysis of the 2nd category data will be open access and available for dissemination to the public (i.e. publications, etc.).

5.2 Protection and management of data

Data types listed in section 3 do not constitute as personal data and their output will be provided at resolutions not relevant for GDPR regulations. Regarding the data collected for and generated from the project's communication

and dissemination activities (Activity 7) personal or any other data will not be collected without prior informed consent of the subjects in questions.

In order to identify, reach, and involve the maximum number of stakeholders, HELCOM have created a dedicated workspace for the project accessible to all partners in the BDF Consortium. This workspace will be also available in the private area of the project's website. In order to prevent malicious use of the workspace information, the following measures have been designed:

- Stakeholders who express an interest in joining the workspace and being involved in the project's activities must be approved by the project coordinator.
- Refrain from gathering personal or any other data without prior informed consent of the subjects in question. If collected, such data will be used exclusively for the purposes of the project and will under no case be shared with third parties.
- Collection of basic institutional data as well as contact details (e.g. occupation/position, organization and e-mail address) upon receiving the informed consent of the stakeholders first. Such data will be used exclusively for the purposes of the project and will under no case be shared with third parties.
- Only publicly available information regarding stakeholders will be collected in case no informed consent is provided.
- The Administrator (i.e the Coordinator of the project) does the full site backup at regular intervals and inspects for intrusions.
- Only authorised users within the consortium can modify data entries.

Moreover, for the purposes of the project, web questionnaires/surveys and relevant forms that will be used asking stakeholders for data relevant to the project's objectives, will commence with an informative section on the project and its data management and protection procedures. This is applicable under Activity 2 of the project.

Furthermore, another form of collecting data (mainly of institutional, scientific, and technical nature) in the framework of the project's communication and dissemination activities is through the organization of physical or online interviews with stakeholders. Such data is expected to feed into the project's activities. In this case as well, the interviewees will be asked for consent.

6. Ethical aspects

According to Article II.5 of the Action Grant Agreement, 'the Agency and the beneficiaries shall preserve the confidentiality of any information and documents, in any form, which are disclosed in writing or orally in relation to the implementation of the Agreement and which are explicitly indicated in writing as confidential'.

Regarding the use of personal data, Article II.6. of the Action GA states that 'Any personal data included in the Agreement shall be processed by the Agency pursuant to Regulation (EU) No 2018/1725. Such data shall be processed by the data controller solely for the purposes of the implementation, management and monitoring of the GA or to protect the financial interests of the EU, including checks, audits and investigations.' HELCOM and BDF partners shall have the right of access, rectify, or erase their own personal data and the right to restrict or, where applicable, the right to data portability or the right to object to data processing in accordance with Regulation (EU) 2018/1725.

HELCOM and project partners must process personal data under the Agreement in compliance with applicable EU and national law on data protection (including authorisations or notification requirements). Article II.6.2 states that the beneficiary may grant its personnel access only to data that is strictly necessary for implementing, managing, and monitoring the agreement. The beneficiary must ensure that the personnel authorised to process personal data has committed itself to confidentiality or is under appropriate statutory obligation of confidentiality.

The beneficiary must adopt appropriate technical and organisational security measures having regard to the risks inherent in the processing and to the nature, scope, context and purposes of processing of the personal data concerned. This is in order to ensure, as appropriate:

- The pseudonymisation and encryption of personal data;

- the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services;

- the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident;

- a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing;

- measures to protect personal data from accidental or unlawful destruction, loss, alteration, unauthorised disclosure of or access to personal data transmitted, stored or otherwise processed.