Document title	Coordinated reporting and hosting of HELCOM continuous noise monitoring data
Code	4-14
Category	DEC
Agenda Item	4 - Matters arising from the HELCOM Groups
Submission date	12.11.2018
Submitted by	Executive Secretary
Reference	Outcome of HOD 54-2018, para. 4.29

Background

HOD 54-2018 noted the open issue of hosting indicator data for continuous noise, emphasized the importance of storing and making the data available, recognized the importance of a potential host to provide continuity and of linking to the similar work done, e.g. under OSPAR in the North Sea. The Meeting requested the Expert Network on Underwater Noise to further clarify what is required for hosting arrangements, and the Secretariat to further explore the possibility and cost of ICES taking up the hosting of the database and soundscape (Outcome of HOD 54-2018, para. 4.29).

Following HOD 54-2018 the EN-Noise agreed to stablish a drafting group with Denmark, Estonia, Germany and Sweden to further clarify what is required for hosting the data and subsequently inform State & Conservation (Memo of the meeting of the EN-Noise held 16 August 2018). The drafting group held an on-line meeting on 23 August 2018 to elaborate a document/inquiry specifying hosting arrangements needs for the continuous noise data, which was subsequently shared with ICES to further explore the possibility and cost of ICES taking up the hosting of the database and soundscape. Based on the input provided by ICES, a revised draft document was shared with the EN-Noise for commenting. Additional comments were provided by Germany and Lithuania. STATE & CONSERVATION 9-2018 considered the revised document on coordinated reporting and hosting of HELCOM continuous noise monitoring data and supported the hosting of the database and soundscape by ICES. The meeting took note that Denmark and Poland would prefer that the database is funded from the HELCOM budget and welcomed that Sweden is willing to co-fund initial cost to set up the database at ICES and that Sweden is of the opinion that the tool should also be hosted there. The meeting took note that Germany is of the opinion that the evaluation of the requirements for soundscape modelling should be addressed as soon as possible, or as soon as the hosting of the data is settled (document 3MA-6 and Outcome of STATE & CONSERVATION 9-2018, para. 3MA.13 - 3MA.16 and 3MA.18).

This document contains information on the technical requirements and estimate cost for (i) the database for the national underwater noise monitoring programs and (ii) the visualization of results for managers, or so called soundscape planning tool. As a summary, the estimated costs as provided by ICES would be as follows:

- for the database, upload and quality control and web services to feed a viewing service: between 10 500 and 13 500 EUR, with an estimated additional yearly cost of maintenance of 5 400 EUR;
- for the soundscape planning tool: in the range of 15 000 EUR (using the existing ICES standard framework as the visualization tool and adapting), plus an estimated yearly cost of maintenance of 400 EUR.

Action requested

The Meeting is invited to take note and consider the information, decide on a hosting solution for a HELCOM database on continuous underwater sound and further deliberate on funding options for hosting indicator data for continuous noise from the Contracting Parties.

Coordinated reporting and hosting of HELCOM continuous noise monitoring data

This inquiry is separated into two: one for monitoring data and one for maps. There is no requirement for them to be included into the same technical solution.

Database for the national underwater noise monitoring programs

The primary aim of the database is to ensure that processed and quality assured data from the national monitoring programs are stored and accessible to HELCOM and thus available for inclusion in HELCOM's recurrent assessments of the environment in the Baltic. A secondary aim of the database is to make the stored data available to HELCOM member states and other public and private organisations with an interest in using the data.

The database is expected to be formed around the existing BIAS[1] data and subsequently amended with national monitoring data on a regular basis.

Format of the input data

Raw data, in the form of continuous or duty cycled sound recordings from stationary monitoring stations, are collected by the HELCOM countries. These recordings are to be quality controlled and processed by the countries, according to the monitoring guidelines and reported to ICES in a data format to be decided. Raw data are stored by the member states in an uncompressed format (wav, or similar).

Data reported to ICES will consist of mean sound pressure levels (dB re. 1 uPa) at regular intervals and in regularly spaced 1/3-octave bands. The averaging time is to be decided, but likely to be between 1 and 20 seconds. BIAS data have a resolution of 20 seconds. The 1/3 octave bands are with centre frequencies as specified by ISO standards (base 10 system), covering the range from 10 Hz up to maximally 160 kHz (i.e. maximally 43 bands), noting that individual datasets may contain data in only a limited number of bands. Data sets may be continuous or regularly duty cycled (e.g. 20 minutes every hour) and may contain shorter or longer periods without data, where for example segments of raw data have been removed during quality control.

Maximum amount of data per station per year (one point every second for 43 bands, continuously) is approximately 5 Gbyte. Amount of data in the BIAS data set is between 10 and 20 Mbyte per station per year (one point every 20 seconds, in three frequency bands, and in many cases duty cycled at 33%). On top of this there will be small amounts of metadata, such as geographical position, sensor information and calibration data. See further below on requirements for metadata. Exact number of stations in the HELCOM area is unknown, but likely to be in the range 20-30 for the foreseeable future.

Specifics about input data

A number of important decisions have to be made on input data. The issues identified so far are listed and discussed briefly below:

- Data file format. The data file format of the BIAS data is ASCII. It is suggested that compatibility with this format is kept for reasons of consistency, but for future amendments it is suggested that a new data format is adapted, such as the HDF5 format described in appendix 1. The new data format should be consistent with the data format adopted by JOMOPANS, which, in turn, is likely to be the data format adopted by OSPAR. A specific standard for data files will be developed at a later point, including specifications for upload of data spanning multiple stations and periods (separate files or one common file).

[1] Baltic Sea information on the Acoustic Soundscape (BIAS) project.

- Mode of upload. Data files are expected to be large, so an efficient way of uploading must be available. In line with the HELCOM/OSPAR impulsive noise register hosted by ICES, it would be advantageous if both an automated and a manual method could be suggested.
  - **Comment from ICES:** There are a number of ways to achieve this, we could explore file compression to ensure we get the smallest packet of data; we can then look at slicing uploads and also whether we are better off doing a background upload (via web services) where the user does not wait online for completion (the last is useful if the source institute does not have a sufficient upload capacity). ICES is also using a corporate DropBox, which may have some advantages if everyone is uploading/downloading at the same time.
- Quality control of data. It is suggested that all quality control of the data remains the responsibility of the data provider. ICES should thus only check received files for consistency and other issues readily identifiable as error. Other issues relating to the quality of the data, whether discovered by, or communicated to ICES, should be forwarded to the data provider and noted in the yearly report from ICES to HELCOM.

The data is naturally organised around deployments. One deployment constitutes one series of data recorded with the same instrument, on the same geographical position, with the same settings and over a specified period of time (starting with deployment and ending with recovery). Variation in data due to systematic and random errors is in general expected to be smaller within deployments than between deployments. For this reason, each data point should be traceable not only to station and instrument, but also to individual deployment. Currently, BIAS data are structured month by month, for each recording station. It is suggested to change this to deployment by deployment, as this will simplify the data file structure, as metadata should be the same within each deployment.

It is to be discussed and decided whether the database should be open for submission of data from other sources than national monitoring programs. This could be noise monitoring programs in connection to large offshore infrastructure projects or research projects. This requires that data is appropriately indexed, so that national monitoring data can be uniquely identified and extracted selectively for analysis. As submitting organisation should be mandatory in the metadata (see below), this indexing is unproblematic. It also requires decisions on who shall pay for the extra workload associated with adding additional data. This could for example be covered by HELCOM member states through a common agreement between HELCOM and ICES or it could be paid for by the individual data suppliers (other than those responsible for the national monitoring). The cost estimate for the hosting of HELCOM continuous noise monitoring data provided in this document does not include cost of potential additional data from such other sources.

Upload routines

Some routines and standards for uploading of data are needed. Countries are requested to upload their quality controlled data on a yearly basis, before some suitable deadline in the year following data collection. After this deadline, ICES should compile and submit a data report to HELCOM, providing information about the amount of data submitted for the previous year. This could be in the form of an online dashboard listing submissions amended to the database, as no processing of the data is required for the reporting.

Database structure

The data are expected to be delivered in the same fundamental format, as described above. Sufficient flexibility must be available to accommodate data from different recording instruments and with different recording regimes. Minimum required information for inclusion in the database is:

- Geographical coordinates (long/lat. WGS84).
- Sequence of sound pressure measurements (Leq) in at least one frequency band. Each value should be expressed as a sound pressure level (dB re. 1 uPa). Each data point should have an associated time stamp (UTC), appropriately corrected for clock drift, if possible.
- Organisation, which supplied the data, for further reference.

Additional information about recording instrument, deployment details, data processing software etc. should be included, in accordance with the HELCOM monitoring guidelines. However, there could be cases where some of this metadata is missing, but where it nevertheless makes sense to allow inclusion of the data, with metadata limited to the minimum requirements. This would typically be the case for historical data.

Extracting data from the database

It should be possible to query the database and download parts of the data for national reporting or other uses.. Such queries should return only the measurements and associated metadata, not extrapolations beyond the measurements, i.e. maps and geographical data layers. Such geographical data should be accessed via the soundscape mapping tool, described further below. Two schemes, not mutually exclusive, are presented as examples. In both cases, some search criteria are specified. This could be combinations of a geographical area, a time period, and/or one or more specified recording stations.

- In the first variant, all available data points fulfilling the search criteria, are exported to one or more export files, which can then be downloaded. Structure of this or these files is to be discussed, but the challenge will be to keep associations between individual data points and their metadata. Alternatively, the user will have to accept some loss of metadata.
- In the second variant, the organisation within deployments is retained. The user could be presented with a selection of deployments fulfilling the search criteria and select which deployments to download. This will greatly simplify the task of keeping track of metadata, as data from each deployment could be exported individually, each with a complete set of metadata attached.

Selection of geographical areas could be implemented in different ways, for example by country, ICES rectangles, ICES fishing zones, HELCOM sub-basins, or just a simple polygon. It may be useful to search also by specific Marine Protected Areas (Natura2000 etc.).

Export data file format

To be discussed and decided. Useful with a selection of options, including a text format for small data sets and a more well-structured binary format for large data files (netCDF, HDF5 etc.).

As data file size could potentially be very large, it is also an issue how the data can be downloaded.

Modelled soundscape data

In addition to monitoring data, it is anticipated that maps of modelled sound in the entire HELCOM area will be generated. There is thus an additional need to store these maps and provide means of accessing them.

Visualization of results for managers

For visualization of the reported data and available sound maps, some kind of interactive online map service or tool is required. One example of such tool is the BIAS Soundscape planning tool, SPT (http://bias.cartesia.se/). In the tool, measured and modeled data can be viewed and own maps and statistics extracted (see the BIAS planning tool user guide). The tool is based on a map platform developed in 2009 - 2014 by Chartiks AB. Many of the platform requirements were formulated by the EU INTEREG Botnia-Atlantica funded project SeaGIS in 2011-2014 and then developed further during the EU LIFE funded project BIAS.

Data storage in the BIAS Soundscape tool is a SQL Server, which thus provides the basis for all data storage in the application. All data access that handles map data is done through the same data storage based on SuperMap objects. The parts of the application that draw map images, find data by location, manipulate data etc. are written in C # and use SuperMap Objects as software libraries. The parts that meet the user in the form of a GUI in the web are written in HTML and Javascript. The parts that are unique to the BIAS tool are written in the same way in HTML and Javascript, but with the addition that google charts are used to display graphics in the web application. It is not a requirement that this organisation of data and maps is retained in transferring data to a new platform. It is possible to break out the BIAS tool from the current application and customize it for another technical platform. There are thus no compelling dependencies to use SuperMap Objects. As long as the same statistical output and maps can be generated as in the existing BIAS Soundscape planning tool, a new platform can be built. In fact, it may be advantageous to change platform, as limitations of the existing platform, such as inability to handle multiple years of data and visualization of other data layers such as animal distribution and habitat characteristics begin to emerge.

Budget estimate

The table below reflects the budget estimate as indicated by ICES. Please note that this information was also submitted to STATE & CONSERVATION 8-2018 (document 3MA-5),.

Name of organisation	International Council for the Exploration of the Sea (ICES) www.ices.dk
Point of contact	Neil Holdsworth, Head of Data and Information neilh@ices.dk
Organization structure	ICES is an inter-governmental organisation, built on a network of more than 5,000 scientists from over 690 marine institutes in 20 member countries. ICES provides scientific advice and technical services to a number of clients including the European Commission, HELCOM, OSPAR, NEAFC and EEA. The ICES data centre manages a vast array of monitoring data, including hosting the Impulsive Noise Register on behalf of OSPAR and HELCOM.
Data sharing platform	
Can you host a database with processed acoustical data that is reported annually (approx. 3-5 GB each year) and environmental input data to the soundscape model	Yes, we host and develop several databases (SQL Server platform) and use virtual machines for scalability in terms of storage and performance scalability.
Can you develop a reporting portal were each nation uploads data and a protocol for quality control? If so, how much time would it take?	Yes, we host and develop several databases (SQL Server platform) and use virtual machines for scalability in terms of storage and performance scalability.
Do you have an example of such database up and running now?	Yes, a few examples: Impulsive noise http://ices.dk/marine-data/data-portals/Pages/underwater-noise.aspx Vulnerable Marine Ecosystems http://ices.dk/marine-data/data-portals/Pages/vulnerable-marine-ecosystems.aspx Acoustic (fisheries) portal http://ices.dk/marine-data/data-portals/Pages/acoustic.aspx
Can you automate the transfer from the database to the BIAS soundscape planning tool?	Yes. Usually we would see this in terms of content architecture; having a data layer (the database), probably a transformation layer (that turns the data in the database into the form that would be ready for use in other applications), a service layer (web services etc. that can be queried by a web application) and the web application (the BIAS soundscape tool).
What is an estimated cost for setting up the above described structure?	For the Data Sharing Platform, upload and quality control and web services to feed a viewing service: Lower estimate = 10500 EUR Upper estimate = 13500 EUR
What is the estimated yearly cost to administrate such database including reporting and quality control?	5400 EUR
In the future, additional soundscape maps or similar and data layers for AIS and VMS data might be produced and needs to be stored as well. Can this future database store data up to 50 GB?	Yes, although it shouldn't be assumed that a relational database is the best means for storing/accessing different data types.
Soundscape planning tool (SPT)	

Do you have the capability to take over the development the SPT, described in the DSP SPT description document? Examples of new features are adding the possibility to have different yearly data series and some restructure of the menus. No new statistical measured are planned for at the moment.	Yes, we have experience with map visualization tools.
What new technical platform do you think this tool could be transferred to for a long term solution?	ICES use both open source and proprietary map software (open layers, geoserver, ESRI/ArcGIS). Depending on the user/client specification, we can adapt or mix between backend/frontend tools as appropriate. In most cases, we recommend to limit customization and the inclusion of 3rd party tools. In addition, we would probably include the calculation scripts and documentation on an open source repository such as Github, in order to facilitate the experts to develop scripts and tools that can either be incorporated or pulled from the database/web services, and ensure a transparent documentation of the tool and methods.
Do you have an example of such visualization tool up and running now?	http://gis.ices.dk/sf/ http://underwaternoise.ices.dk/map.aspx
What is an estimated cost for setting up the above described structure?	In the range of 15000 EUR, however this is based on a few assumptions (using the existing ICES standard framework as the visualization tool and adapting).
What is the estimated yearly cost to administrate visualization tool including automated input of data from the data sharing platform?	400 EUR.

Appendix 1

Hosting requirements and suggestions for an exchange-format for underwater sound monitoring data

Based on experience gained from the German national Monitoring program of underwater sound, this document reports definitions and requirements for hosting and handling continuous underwater sound data.

Hosting continuous sound data requires storage and organization of large amounts of data. The data formats HDF5 and NetCDF are both well suited for this purpose and technically widely supported. Due to high level of compatibility between both formats and the abundance of technical support for conversion, an explicit definition for the suggested exchange-format considers HDF5 data structure.

The HDF5 format provides objects called *groups*, *datasets* and *attributes*. A *group* is comparable to a folder in a file system. *Datasets* can be e.g. matrices or single values/strings. *Attributes* can be used to store metadata of datasets. Resources in HDF5 files can be accessed using a POSIX-like syntax e.g. */filename/group/specific_resource*.

For a lean exchange format definition, the hdf5-exchange-format can be composed by *groups* and *datasets* only:

- Group: a container structures composed of datasets and other groups
- Dataset: single value/multidimensional arrays of a homogeneous type

For more information on the hdf5 format and supporting software follow the link www.hdfgroup.org.

Explicit suggestion for an exchange-format definition

The following format definition describes the suggested hierarchy in the HDF-file and specifies the suggested datatype (e.g. int, float, string and bool).
All **bold** names followed by a <HDF dataset …> are *datasets*. All names followed by further **names** are *groups*. The highest parent node is the hdf-file in the file system "result_file.h5":

```
result_file.h5/
    /dataset_ambient_noise
        # general metadata for monitoring data
        /name_measurement_project # data organization/assignment
            <HDF dataset, type string>
        /name_measurement_position # data organization/assignment
            <HDF dataset, type string>
        /measurement_system_id  # technical info: unique
            <HDF dataset, {'MS-CC-1234'}, type string>
        /measurement_id  # technical info: unique
            <HDF dataset, shape (), type string>
        /cordinates_measurement_position # station coordinates
            <HDF dataset, shape (2, ), type float
             (latitude, longitude) in Decimal degrees WGS84, 7 decimals>
        /measurement_height  # in meters, height above ground
            <HDF dataset, shape (), type float>
        /measurement_purpose
            <HDF dataset, {e.g. 'Monitoring'},
```

```
     type string>
/evaluation_type_leq
    <HDF dataset, {‘pulse’, ‘block’}, type string>
/hydrophone_type
    <HDF dataset, {e.g.: ‘TC-4033’}, type string>
/hydrophone_serial_number
    <HDF dataset, type string>
/calibration_level  # converts rawdata levels to sound pressure levels.
    <HDF dataset, type float>
/system_calibration_procedure
    <HDF dataset, {‘electric’, ‘acoustic’}, type string>
/rawdata_uuid
    <HDF dataset,
    {example: ‘e9b82f55-38f2-48a5-bc76-2cb651dd285f’}, type string>
/comments
    <HDF dataset, type string>
/dataset_version
    <HDF dataset, {example: ‘1.0’}, type string>
/dataset_type
    <HDF dataset, {‘ambient_noise’},
     type string>
# monitoring data part
/frequency_count # number of frequencies
    <HDF dataset, shape (), type int>
/frequency_index
    <HDF dataset, shape (frequency_count, ), type float>
/leq_count
    <HDF dataset, shape (), type int>
/leq_temporal_values
    <HDF dataset, shape (leq_count, ), type float>
/leq_spectro_temporal_values
    <HDF dataset, shape (), type float>
/leq_averaging_time
    <HDF dataset, shape (), type float>
/leq_datetime_index
    <HDF dataset, shape (), type int>
```

The storage of datasets containing time series of continuous data requires special attention. All related datetime vectors should be stored as POSIX timestamps in UTC. This ensures an easy data handling with every programming language. Avoiding an overflow problem in the year 2029, they should to be stored as int64 type.
All other datasets can be expected to vary within comparably limited value ranges and should be stored as 32 bit data types. It is recommended to allow for a dynamic adaption of exchange data accuracy.

HDF5 file size in dependence on temporal and spectral sampling

The choice of a common spectral and temporal sampling of monitoring data depends on the aspired further use of data. Monitoring data HDF5 file sizes for storing 1.5 month of continuous data using the format suggested above are approximated for a) a minimal and b) an extended spectral and temporal sampling scenario. Using the HDF5 storage layout properties ‘contiguous’ and ‘chunked’, estimated file sizes yield

a) 3 frequency bins and L_{eq} averaging time of 20 sec: approximately 2 MB
b) 33 frequency bins and L_{eq} averaging time of 5 sec: approximately 200 MB