GEANS Data Management Plan V1.2

GEANS

Genetic Tools for Ecosystem Assessment in the North Sea Region

CONTENT

GENERAL INFORMATION

Version	1.2
Version date	21/10/2022
Reason for change	Inclusion of project extension activities, like a new WP on eDNA + general refreshment

INTRODUCTION

Benthic species diversity and abundance are common indicators for ecosystem health. Therefore, monitoring of benthos is important in order to detect ecosystem changes and apply effective management. However, the traditional techniques used to monitor benthic species are often time consuming, labor- and cost-intensive, and require taxonomic expertise for species identification. In contrast to these traditional methods, Genetic tools for Ecosystem health Assessment in the North Sea region (GEANS), an Interreg-North Sea region project, will use DNA-based methods to assess ecosystem health. We aim to reduce costs and time for benthic monitoring, while increasing accuracy. DNA-based methods do not require taxonomic expertise and methods can be more easily standardized into standard operating procedures (SOPs). Since DNA-based ecosystem assessment is an upcoming field of research, harmonized protocols for monitoring marine ecosystem health are still under establishment. To this end, seven North Sea region partner institutes collaborating in the GEANS project will (i) develop and standardize genetic monitoring tools for ecosystem health assessment, (ii) apply these tools and SOPs in pilot studies to retrieve biotic indicators and (iii) build a decision framework to support the implementation of the standardized tools to facilitate ecosystem health assessment and management decisions in the North Sea region. The pilot studies take a central role in the GEANS project as they prove the advantages of the applied tools while simultaneously ensuring engagement and exchange of knowledge between stakeholders and scientists. So far, three pilot studies have been initiated: soft-bottom benthos monitoring, hard bottom substrate monitoring using Autonomous Reef Monitoring Structures (ARMS) and detection of non-indigenous species (NIS) in harbors.
Over the course of the project, several data types will be generated by the GEANS partners and these will vary strongly among the project work packages (WPs). The project consists of six Work Packages (WP); WP1 - Project management, WP2 - Communication, WP3 - DNA reference library, WP4 - Harmonisation and consolidation of genetic tools and protocols, WP5 - Pilot studies and WP6 – Translating science into products, WP7 - Zero-impact monitoring for ecosystem health and restoration (eDNA).

The purpose of this data management plan is to set up a structured plan to make this data FAIR and thus Findable, Accessible, Interoperable and Re-usable (Wilkinson et al., 2016) both during the project as well as after it ends. This data management plan will give an overview and detailed description of the data types that will be generated, how these data will be collected, the platforms where the data will be stored and how the data will be accessible and preserved per WP.
WP1 and WP2 generate data throughout the project and are therefore also described in this plan. However, the focus of this DMP will be on research data and therefore, the measures for working towards FAIR data only apply to WPs 2, 3, 4, 5 and 7.

DATA SUMMARY

WP1: PROJECT MANAGEMENT

During the project, partners will have to measure up to certain progress and financial reporting. Every partner will have to report their progress for each WP and submit their financial claims. This data will be mainly textual and numbers on the output indicators and on targets set for different activities. Internal finance data will be reported in excel sheets. The data formats will be .PDF and .XLSX. The total amount of documents at the end of the project will be around thirty requiring a storage of only a few MBs.

All management data from WP1 will be collected and stored on the Interreg Online Monitoring System (OMS). The project management will set deadlines for progress and financial report submission, project partners can log on OMS and submit their documents. The progress reports will be checked by Interreg JS, but also by the partners. The progress reports are then compiled in an overall project report. Financial claims will be checked by the FLC (First Level Control), these financial reports are also available through the OMS system.

WP2: COMMUNICATION

WP2 will result in multiple data types used to communicate through different channels such as project websites (https://northsearegion.eu/geans and https://www.geans.eu/) and social media (https://twitter.com/GEANS_Interreg). The data displayed on the output website geans.eu will be mainly textual (presentations, scientific papers, reports, press releases, newsletters, protocols, fact sheets and stories and posters) and visual (pictures, logos and video footage). Depending on final video data, total storage space will be several GBs. In addition, an overview of website traffic data will be collected as well in .XLSX or .PDF format over the course of the project. Content displayed on the geans.eu output website will be archived in the Marine Data Archive (MDA).

WP3: DNA REFERENCE LIBRARY

For WP3, a DNA sequence reference library will be constructed for North Sea benthic fauna, including both quality-checked existing sequences and newly generated sequences. All the sequences will be accompanied by the relevant collection and biological metadata. In the end the reference library will be built up by three types of reference data:

I. GEANS core species reference collection: GEANS generated sequences
II. GEANS partner species reference collection: sequences generated by GEANS partners prior to the project
III. BOLD/GenBank background species reference collection: sequences generated by non-GEANS partners used in GEANS

In a first step, a key species list for the North Sea area will be created, including non-indigenous species. Existing in-house and public reference databases will be scanned for DNA sequences of

the marker genes COI, 18S, 16s and 28S for these species. In-situ reference specimens will be collected from the North Sea to be sequenced using Oxford Nanopore MinION (Jain et al., 2016). The raw sequence data will be quality filtered, trimmed and assembled while trace files will be produced. To process the genetic data, the bioinformatics software Geneious (Kearse et al., 2012) will be used. In this way, the quality characteristics (peaks, phred scores, etc.) of each sequence will be checked. The sequences will then be control-checked using BLAST (Basic Local Alignment Search Tool), to see if it does indeed represent the voucher specimen. For each species two specimens will be sequenced. In case the sequences are not matching, a third specimen will be sequenced. Finally, taxonomic experts will be advised in case of species complexes or nomenclature issues and museum collection databases will be advised in case of taxonomic issues. A photo library of pictures of specimens used for sequencing will be created using Adobe Lightroom for an efficient management and editing of the photos. Finally, all the produced genetic data, photos, trace files will be uploaded in BOLD (Ratnasingham & Hebert, 2007), a free and public database where the user can process and analyse the data produced.
The reference sequences will be accompanied with appropriate metadata including information on taxonomic coverage (i.e. AphiaID from the World Register of Marine Species (WoRMS)(Costello et al., 2013)) and data on the origin of the sequence (i.e. Collection information: coordinates, date, etc.; Biological information: juvenile, adult, female, male, etc.). For each specimen a scaled image will be uploaded to give information on morphological characteristics and size. Tabular and textual metadata will be delivered as well, containing both collection and biological information. The expected data formats will be .FASTQ (raw sequences) and .FASTA (processed sequences) for the sequences and ABI for the according trace files. The picture formats will be .TIFF/.PNG/.JPG. The metadata will be delivered as .PDF/.XLSX. Primers used to produce the sequence for each species and will also be included in this dataset. Similar to the species reference data, the files will be added in .FASTA and will require information on the author, data provider, and date.

The estimated size of data for WP3 will be tens of GBs. At least a thousand photos are estimated to be produced, each with a data size of around 5 MB. Also, hundreds of trace files (~265Kb each) will be produced. A list of key species will be produced including 800 rows of species, accompanied by metadata in about 20 columns.

WP4: HARMONIZATION AND CONSOLIDATION OF GENETIC TOOLS AND PROTOCOLS

During the project, methods used for metabarcoding will be evaluated to set-up harmonized standard operating procedures. The study will be based on literature research, project questionnaires and evaluations of several pilot studies. The collected data will be textual data formatted as .PDF. In total, the size of the data will be in the range of MBs. This data will be archived in the Marine Data Archive (MDA).

WP5: PILOT STUDIES, WP7: ZERO-IMPACT MONITORING

Through pilot studies, protocols will be tested to produce a set of Standard Operation Protocols (SOPs) that will be recommended and applied by all partners. The data will originate from both long-term monitoring and new field sampling. Within these pilots, three major pilot types can be distinguished: the non-indigenous species pilot (NIS), the soft sediment pilot (SBS), the hard

substrate pilot (HBS) for which the ARMS network will be used. In the GEANS extension, a pilot study on eDNA under WP7. Each partner will carry out one or more of these pilots to monitor macrobenthos. From the samples, DNA will be isolated and metabarcoded. For the NIS, SBS and eDNA pilots, the generated data will be sampling and environmental metadata (.XLSX/.CSV), traditional species observation data (.CSV/.XLSX/.PDF) and according images (.TIFF/.PNG/.JPG). Raw sequence data (.FASTQ), post-processing sequence data (.FASTA), clustered/OTU table or ASV files (.CSV), taxonomically assigned OTU or ASV tables (.CSV) will be uploaded in case scripts and parameter settings are not shared. In case scripts and parameter settings are shared, input data (raw sequence data (.FASTQ)) and output data (taxonomically assigned ASV or OTU table) will be provided. Traditional and genetic methods will be compared in terms of time and cost efficiency, these calculations will be available in .XLSX format. The data size is expected to be tens of GB's per pilot study. This data will be archived in the Marine Data Archive (MDA), the European nucleotide Archive (ENA) or Barcode of Life Datasystems (BoLD), and linked to the IMIS metadata discovery record for the dataset. For the hard-bottom substrate pilot, data will be processed and archived according the ARMS protocols and dataflows[1].

WP6: TRANSLATING SCIENCE INTO PRODUCTS

To inform stakeholders, and get their feedback on GEANS recommendations, several documents will be generated. ICES TIMES documents and policy papers will be .PDF format. Reports of workshops and meetings between partner institutes and stakeholders, will be generated in .PDF format. This data will be archived in the Marine Data Archive (MDA). To guide stakeholders and users to appropriate protocols, pilot studies, papers, factsheets and stories, a decision support structure in the form of a tree or network will be hosted on the GEANS output website at geans.eu. An interactive export of the framewerk will be archived in the Marine Data Archive (MDA).

MAKING DATA FINDABLE (F)

Data generated by GEANS are findable through discovery metadata in the Integrated Marine Information System (IMIS). The metadata will be structured by assigning 'parent' datasets and child datasets for pilot studies, reference library and zero-impact monitoring. The parent dataset is linked to a main project dataset. Each IMIS record has a unique URI, which can be used as an identifier. Contributors of partner institutes will be asked if they want a personal IMIS record created, which will be linked to datasets they contributed to. Papers published under GEANS will be uploaded to IMIS, and records can then be linked to dataset records.

To facilitate proper data management, partners will be provided with a template spreadsheet in which they can fill in their dataset metadata to be entered in IMIS. A list of what needs to be provide includes:

- A title which covers the content of the dataset
- The person who is responsible for the dataset

[1] https://github.com/arms-mbon/

- A description of the dataset. For each dataset the abstract is included in the metadata template
- A citation of the dataset, how it should be cited
- People who contributed to the dataset
- Measurements
- Coverage (Spatial, Temporal & Taxonomical)
- License
- Link to access the data

Repositories holding the research data itself will be linked to their metadata records in IMIS. Depending on the work package, data will be stored on a specific database. All WP3 reference library data will be submitted to BOLD. A private BOLD project will be created to store all of the DNA reference data. This private project will then be made public at the end of the GEANS project. This includes the raw sequences, metabarcoding sequences, environmental data, images and trace files. Data from the hard bottom substrate pilot are stored according to ARMS data-management; raw data files on the European Nucleotide Archive (ENA), ENA accession numbers, FASTA files, OTU and ASV tables, metadata and image data can be found on PlutoF. For most recent decisions on dataflow and repositories for ARMS data, we refer you to the ARMS DMP available at https://github.com/arms-mbon. All data generated by WP4, WP5 (NIS & SBS) and WP7 will be archived on the Marine Data Archive (MDA), a trusted and open-access repository. To manage data availability, a public as well as a private (shared among the project partners) folder will be created in MDA. For instance, the data will be archived in the private folder and after a moratorium period, data will be moved to the public folder and become openly available. By the end of the project, all of the data should be moved towards the public folder.
In order to access the shared folder, users must register on https://mda.vliz.be/account.php and create a free account. After registration, users will be able to access the shared GEANS folder. Folders in MDA are structured according to WP, with sub structure according to institute for WP5 pilot studies and WP7 or type of output for WP2

Since all data will eventually be open data, Digital Object Identifiers (DOI's) will be assigned to them. A DOI is a permanent identifier, data (versions) should be static after submission for an identifier. MDA will make sure the data remains persistently available. Within publications, a DOI will be added to allow easy access to the data.

MAKING DATA ACCESSIBLE (A)

Since there is a grant agreement on open data, all research data produced by GEANS should be open by the end of the project. Pending publication, the data is temporarily placed under moratorium (MO unrestricted after moratorium). As soon as the data is published, this data is assigned the creative commons license. By the end of the project duration all data is assigned the creative commons license (CC-BY-4). Under this license, users are allowed to copy and redistribute data, adapt data as long as appropriate credit is given and changes are indicated. Papers published under GEANS will be made openly available through IMIS under the Open Access provision in Belgian Law, regardless of journal policies.

A new GEANS output website[2] will serve as a central gateway to access the databases/metadata records/protocols/project/publications/presentations/posters/factsheets and stories. Within this output website, a decision support tool will be generated under WP6 to guide users towards the data they are searching for. Even more, usage of international systems and data flows will be maximally encouraged within the consortium (e.g. EurOBIS, WoRMS). Partner institutes will be supported to archive their data in European systems like EurOBIS in the appropriate formats. In this way data can be properly managed, prior to archiving the data and data flow towards international systems.

MAKING DATA INTEROPERABLE (I)

By working towards standard operating protocols, methods and software will be maximally harmonized between partners. Bioinformatics harmonisation will be discussed continuously during the project due to continuous evolution of relevant tools.
Metadata are made interoperably by adhering to globally accepted metadata standards. The IMIS datasets catalogue uses the EML metadata schema (a standard for ecological data) as well as .JSON and .XML. IMIS uses a number of standards in its fields, like the World Register of Marine Species (WoRMS for taxonomy; Marine Regions and FAO ASFA geoterms for geographic locations and for thematic keywords. Concerning the data, BOLD uses DarwinCore standard vocabulary for data delivery. For biological and environmental data, it is advised to appeal to OBIS-ENV-DATA standards, based on the DarwinCore standards https://bdj.pensoft.net/articles.php?id=10989). For delivery of taxonomic data in MDA, it is also strongly advised to use DarwinCore as a standard (http://rs.tdwg.org/dwc/). When submitting files to MDA, one should take the following criteria into account for naming files: Title that covers the content, avoid use of spaces and special characters, avoid long titles, keep it brief, avoid single files, and make use of zip files.

MAKING DATA REUSABLE (R)

To ensure re-usable data, research data will have a CC-BY-4 license allowing a wide re-use of the data without restriction. Under this license users are allowed to share, copy and redistribute the material in any medium or format, and adapt, remix, transform, and build upon the material, the data for any purpose, even commercially. User must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

Data will remain re-usable after the end of the project by archiving in long-term repositories, like BOLD, ENA and MDA.

During GEANS data collection, various stages of quality control will take place, depending on the WP and data type. Data will be initially controlled by the institute that delivers the data. Along the data life cycle, additional quality controls will happen. Metadata quality control will take place during the input of

[2] https://www.geans.eu/

metadata in IMIS. Quality of the delivered data and metadata is the primary responsibility of the owner/creator of the data/metadata.

RESPONSIBILITIES AND SECURITY

VLIZ will proactively support the data delivery and upload by partners of WP4, WP5 and WP7. The final responsibility for implementing the DMP at a local level lies with the partners. There will be central efforts to support the partners by organizing the necessary training and guidelines. Filled-in metadata templates delivered by the partners will be uploaded in IMIS by VLIZ. VLIZ also supports the development of the output website with the integrated decision support framework that will both be used to access the data.

FUNDING

This research was supported as part of GEANS, an Interreg project supported by the North Sea Programme of the European Regional Development Fund of the European Union

REFERENCES

Costello, M. J., Bouchet, P., Boxshall, G., Fauchald, K., Gordon, D., Hoeksema, B. W., Poore, G. C. B., van Soest, R. W. M., Stöhr, S., & Walter, T. C. (2013). Global coordination and standardisation in marine biodiversity through the World Register of Marine Species (WoRMS) and related databases. PloS One, 8(1).

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology, 17(1), 239.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., & Duran, C. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28(12), 1647–1649.

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). Molecular Ecology Notes, 7(3), 355–364.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., & Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3.

https://onderzoektips.ugent.be/en/tips/00001734/#:~:text=The%20law%20allows%20the%20author,a%20clause%20is%20not%20enforceable, accessed 21/10/2022.