

Project Number: BiodivClim.421

Project Acronym: GenClim

Project title:

Biodiversity on the run: evolutionary and socioeconomic consequences of shifting
distribution ranges in commercially exploited marine fishes

DATA MANAGEMENT PLAN

Executive summary. The purpose of the current document is to present the 1st Data Management Plan (DMP) of the **GENCLIM** project. The document has been compiled with the collaborative work among the coordinator and the consortium partners who are involved in data collection, production and processing. It includes descriptions of the types of datasets that will be collected, processed or generated in all Work Packages (WP) during the course of the project duration. The DMP is a living document to be updated as the implementation of the project progresses and when significant changes occur.

Introduction. The present DMP represents the first version of the DMP of the **GENCLIM** project. The current document has the purpose to ensure proper and sound management of the research data that will be collected, processed and generated within GENCLIM. The DMP describes the data management life cycle for the data to be collected, processed and/or generated by **GENCLIM** project, as a BiodivERsA project. The DMP aims at defining the management strategy of data generated during the project with the purpose to making research data findable, accessible, interoperable and re-usable (FAIR). The DMP is not a fixed document, but it is likely to evolve during the whole lifespan of the project. The upcoming versions of the **GENCLIM** DMP will have a clear version number and include a timetable for any occurring data updates.

Data Managers. GenClim has two dedicated Data Managers (DM): the coordinator Romina Henriques (DTU), who will oversee the implementation of the DMP, and Rita Castilho (CCMAR) who will liaise with WP and task leaders for data management.

1. Data Summary

The **GENCLIM** DMP aims to provide a strategy for managing key data generated and collected during the project and optimize access to and re-use of research data. The DMP is intended to be a ‘living’ document that will outline how the **GENCLIM** research data will be handled during and after the project, and so it will be reviewed and updated at regular intervals. In this regard, the main purpose of the DMP is to ensure the accessibility and intelligibility of the data generated during the **GENCLIM** project, where “data” is defined as the full range of digital outputs of the project, from primary binary data, to research software, protocols and text-based policy briefs. All outputs will be subject to FAIR Principles (not just experimental data) to ensure independent validation, reproducibility, and optimal re-use of **GENCLIM** method and outputs.

Data types will fall in four main categories: i) field samples and their metadata (physical tissue samples of the species, including information in .csv files regarding their date and location of collection, size and sex of the individual); ii) genomic data (raw .fastq files, as well as .vcf files containing the final genomic datasets); iii) forecasting model data (.csv files)

and iv) socio-economic models (code in .txt files), in order to achieve the objectives of the project. More specifically, data outputs and types include:

Output: *Field Data*

Type: *Primary Datasets*

Procedure: Field sampling datasets will be archived in [Pangaea.de](https://pangaea.de), fish genomics sequences in [Genbank](https://www.ncbi.nlm.nih.gov/genbank/), with full open access at generation where possible. Physical sample collections, held at consortium partners, with full metadata, to allow re-analysis post-project.

Partner responsible: ISPA

Output: *Environment-linked genotypes Products*

Type: *Secondary Data*

Procedure: Environmentally-linked genotypes will be a DOI-stamped product stored in the GitHub VRE, in .csv format(s) that optimise its re-use for parameterisation of any forecasting model beyond the one developed by the consortium.

Partner responsible: DTU

Output: *Protocols & Methods (GWS and GT-Seq) Methods*

Type: *Protocols /*

Procedure: experimental protocols for genome wide sequencing and genotyping via GT-Seq will be archived in Zenodo.org in parallel to any possible peer-reviewed publications in method journals

Partner responsible: DTU

Output: *Map of Genomic Vulnerabilities Maps*

Type: *GeoSpatial*

Procedure: Mature map products will follow interoperability formats of end-users ICES and EMODNet compliant with EU's INSPIRE Directive.

Partner responsible: DTU

Output: *Short- & Long-term Forecast Model / Bioeconomic Algorhythm Code*

Type: *Research*

Procedure: Model code, description and output simulations archived with DOI in GitHub VRE, access to project partners until validation and publication, full public access after

Partner responsible: KU and DTU

Output: *Peer-Reviewed*

Type: *Research Publications*

Procedure: *Green Open Access* to *pre-prints* and *post-prints* archived at Zenodo.org (GenClim Project Collection) will allow 100% Open Access to all our publications.

Partner responsible: Lead Authors

Output: *Policy Briefs Text*

Type: *Grey Literature,*

Procedure: DOI-issued objects archived in Zenodo.org collection and disseminated according to advice from end-users ICES and STECF Experts within consortium.

Partner responsible: Lead Authors

Output: *Teaching Materials*
Plan(s)

Type: *Lesson*

Procedure: Publications in journal “[Frontiers Young Minds](#)” have DOI by default. Lessons plan(s) co-created by GenCLIM and high-school teachers will have DOI’s issued within Zenodo.org Project Collection and disseminated by [Int. Baccalaureat](#) high-school teachers

Partner responsible: DTU

GENCLIM commits to full reproducibility and will generate all outputs in a Virtual Research Environment (VRE). All data sets generated will be stored in each of the participant entities databases (for long-term storage) and in the shared cloud of the University of Kiel (<https://cloud.rz.uni-kiel.de/>) created as an internal database of the partners. This will be curated by each partner in location, and overall by the coordinator of the project (DTU). Within the shared database in the cloud, data will be organized per work package: WP 2: sample information; WP 3: genomic data; WP4: forecasting models, and WP5: socio-economic models.

Each data set created during the project will be assessed and categorized as open, embargoed or restricted by the owners of the content of the data set. In addition, those categorized as open or embargoed will be publicly shared (in the case of embargoed, after the embargo period is over) through the public section of the project website, the project Github page (<https://github.com/GenClim>) and the project ZENODO collection (<https://zenodo.org/communities/genclim>). All partners will have access to Github and Zenodo project areas.

The Github platform will be used mostly to host the code used in the project, through three main repositories: WP3: genomics (DTU, ISPA, CCMAR and SU), WP4: forecasting (DTU), WP5: socio-economic modelling (KU). Mature versions of the code will be uploaded by the responsible partners. Intermediate versions of the code will be hosted in each partner repository.

GENCLIM data management strategy

This document addresses for each data set collected, processed and/or generated in the project the following elements: Dataset reference and name; Internal project Identifier for the data set to be produced. This will follow the format:

ProjectName_WorkPackageNumber_TaskNumber_PartnerName_DataSubset_DatasetName_Version__DateOfStorage, where the Project Name is **GENCLIM** and the Partner Name represents the name of the data custodian (WP Lead/ Task Leader).

A central database will be created by the coordinator of the project (DTU) containing all the data types and outputs generated in the project in the shared cloud platform. This will be maintained by the WP leaders, and will include the following information:

Dataset description: description of the data generated or collected, including its origin, nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the potential for integration and reuse.

Standards and metadata: reference to existing suitable standards. If these do not exist, an outline on how and what metadata will be created.

Data sharing: description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling reuse, and definition of whether access will be open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating the type of repository (institutional, standard repository for the discipline, etc.). In cases where the dataset cannot be shared, the reasons for this will be stated (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

Archiving and preservation (including storage and backup): description of the procedures to be put in place for long-term preservation of the data, including an indication of how long the data should be preserved, the approximate end volume, associated costs, and how these are planned to be covered.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

GENCLIM open data will be curated in a dedicated ZENODO collection, as its repository structure, facilities and management are in compliance with FAIR data principles. ZENODO is an OpenAIRE repository that allows researchers to deposit both publications and data, providing tools to linking them to these through persistent identifiers and data citations.

All mature outputs will be issued a DOI, and archived in ZENODO Collection at creation, with Creative Commons CC licences for re-use. **GENCLIM** ZENODO Collection will be 100% open at project end, or before (only exceptions applicable will be PhD thesis in progress) and all outputs linked via metadata. **GENCLIM** will adhere to principles of removing all barriers for academic and commercial sectors to create derivative products and services, as long as **GENCLIM** DOI's are cited and credited.

Metadata is data on the research data themselves. It enables other researchers to find data in an online repository and is, as such, essential for the reusability of the dataset. By adding rich and detailed metadata, other researchers can better determine whether the dataset is relevant and useful for their own research. Metadata (type of data, location, etc.) will be uploaded in a standardized form. This metadata will be kept separate from the original raw research data.

The bibliographic metadata will include all of the following:

- the name of the action, acronym and grant number;
- the publication date, and length of embargo period if applicable
- a persistent identifier

Search keywords

Zenodo allows to perform simple and advanced search queries on Zenodo using the keywords. Zenodo also provides a user guide with easy to understand examples.

Naming conventions

Files and folders at data repositories will be versioned and structured by using a name convention consisting as follow: GENCLIM _Dx.y_YYYYMMDD_Vzz.doc

Version numbers

Individual file names will contain version numbers that will be incremented at each revision (Vzz).

2.2. Making data openly accessible

In order to maximise the impact of **GENCLIM** research data, the results will be shared within and beyond the consortium. Selected data and results will be shared with the scientific community and other stakeholders through publications in scientific journals and presentations at conferences, as well as through open access data repositories.

The **GENCLIM** project datasets are first stored and organized in a database by the data owners (personal computer, or on the institutional secure server) and on the project database (KU cloud). All data are made available for verification and re-use initially by the partners, and as mature data in online repositories, unless the task leader can justify why data cannot be made openly accessible. To protect the copyright of the project knowledge, Creative Commons license will be used. The **GENCLIM** dataset deliverables will be made public (data access policy unrestricted) and they will be accessible by:

- project website (the project website will be used as a repository for project deliverables, and results, as well as linking to the Zenodo Collection and Github repositories)
- Partners database
- ZENODO and Github platforms
- Journals

All mature data deposited on ZENODO will be accessible without restriction for public. For other data, potential users must contact the team or the data owner in order to gain access. If necessary, appropriate procedures (such as non-disclosure agreements) will be used.

2.3. Making data interoperable

Partners will observe OpenAIRE guidelines for online interoperability, including OpenAIRE Guidelines for Literature Repositories, OpenAIRE Guidelines for Data Archives. These guidelines can be found at: <https://guidelines.openaire.eu/en/latest/>. Partners will also ensure that **GENCLIM** data observes FAIR data principles under H2020 open-access policy: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-datamgt_en.pdf

Information relating to the interoperability **GENCLIM** datasets will be collated in a central database (see Section 1). As the project progresses and data is identified and collected, further information on making data interoperable will be outlined in subsequent versions of the DMP. In specific, information on data and metadata vocabularies, standards or methodology to follow to facilitate interoperability and whether the project uses standard vocabulary for all data types present to allow interdisciplinary interoperability.

2.4. Increase data re-use (through clarifying licences)

Data will be issued with a Creative Commons CC licences for re-use, and to protect the ownership of the datasets. Both Share-Alike and NonCommercial-ShareAlike licenses will be considered for the parts of datasets for which the decision of making that part public has been made by the Consortium.

However, an embargo period may be applied if the data (or parts of data) are used in published articles in “Green” open access journals. The recommended maximum embargo period length by European Commission is 6 months.

For datasets deposited on a public data repository (ZENODO) the access is unlimited.

An internal process of quality evaluation of data will be implemented throughout the entire project duration to assess both project data /products and project process. An internal peer review will be performed for the main project deliverables to guarantee the deliverables are developed with a high level of quality. Each WP leader has to submit all the produced documents to another partner assigned as internal reviewer to check for the quality of the documents produced.

Once the project data is submitted to curated databses, such as Genbank, the data will remain re-usable indefinitely.

3. Allocation of resources

The costs for making data FAIR includes:

- Fees associated with the publication of scientific articles containing project’s research data in “Gold” Open access journals. The cost sharing, in case of multiple authors, shall be decided among the authors on a case-by-case basis.
- Project Website operation: to be determined
- Data archiving at ZENODO and on other on line data base: free of charge
- Copyright licensing with Creative Commons: free of charge

Each partner is responsible for the data they produce. Any fee incurred for Open Access through scientific publication of the data will be the responsibility of the data owner (authors) partner(s).

4. Data security

All research data underpinning publications will be made available for verification and re-use unless there are justified reasons for keeping specific datasets confidential. The main elements when considering confidentiality of datasets are:

- Protection of intellectual property regarding new processes, products and technologies where the data could be used to derive sensitive information that would impact the competitive advantage of the consortium or its members,
- Commercial agreements as part of the procurements of components or materials that might foresee the confidentiality of data
- Personal data that might have been collected in the project where sharing them is not allowed by the national and European legislation.

The following guidelines will be followed in order to ensure the security of the data:

- Store data in at least two separate locations to avoid loss of data;
- Encrypt data if it is deemed necessary by the participating researchers;
- Limit the use of USB flash drives.
- Label files in a systematically structured way in order to ensure the coherence of the final dataset.

All project deliverables and data will be stored and shared in the Kiel University cloud area allocated to the project. As an initial step, only the Consortium Partners will have access to the cloud storage where dataset and metadata are filed. Following scientific publications and articles, the dataset deliverables and the final demonstrator research results will be shared through ZENODO and other databases.