BIOPOLIS

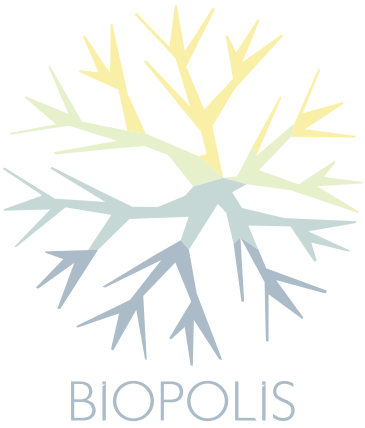
WP4

Data Management Plan including the Open Research Data Pilot, ORDP

Deliverable 4.7

cibio

Porto Business School



Data Management Plan including the Open Research Data Pilot, ORDP

Deliverable 4.7

Lead beneficiary	CIBIO/InBIO
Submission date	May 31st 2020

Table of Contents

1. DATA MANAGEMENT PLAN OVERVIEW

The EC funded **BIOPOLIS project** supports the upgrade of the research unit of ICETA, CIBIO, to a Centre of Excellence in the areas of Environmental Biology, Ecosystem Research and Agrobiodiversity, through extensive Teaming activities with the partner University of Montpellier UM), France, and the Porto Business School (PBS), a business partner . The main goals of the project are to pursue activities within the scope of research and innovation, knowledge transfer and address pressing societal challenges through three main strategic pillars: 1) Environment and Biodiversity assessment and Monitoring; 2) Ecosystem Function and Restoration; and 3) Agrobiodiversity, Conservation and Competitiveness of local Genetic Resources and Farming systems. Through these three main strategic pillars and during the execution of the BIOPOLIS project a large amount of different data types will be generated, managed and disseminated. This document represents the **Data Management Plan (DMP)** for the BIOPOLIS project. The DPM comprises information on what types of data that will be generated within the scope of the project, how it will be exploited and stored, as well as made accessible for verification and re-use, assuring reproducibility and quality of the data. The underlying principles for this DMP are that the data should be managed so that it is findable, accessible, interoperable and reusable (FAIR), according to the *Guidelines on FAIR Data Management in Horizon 2020*. The DMP takes extensive advantage of the expertise acquired from the previously EC funded ERA Chair in Environmental Metagenomics (EnvMetaGen) gained at CIBIO (Project number *668981*), and exhibits strong synergies with the DPM established for the recently initiated, and EC funded, ERA Chair project in Tropical Biodiversity – TROPIBIO (Project number *854248*).

In addition, the DMP was also developed in accordance with PORBIOTA – Portuguese e-Infrastructure for Information and Research on Biodiversity (http://www.porbiota.pt/en), coordinated by CIBIO, and is the Portuguese node of LifeWatch ERIC (https://www.lifewatch.eu/).

This document represents the first version (v1) of the BIOPOLIS DMP and it will, according to the Grant Agreement (GA) of the project, be further updated and developed following the recruitment of the Directors and unit Heads, during the course of the project (Versions 2 and 3 are due at M42 and 78, respectively), and/or at more frequent intervals if considered

necessary by the dedicated Data Management and Open Access Committee (additional information below under sections 5 and 6, and the GA under WP5, Task 5.5).

2. TYPE OF DATA

The BIOPOLIS project will generate large amounts of data related to Georeferenced Biodiversity, Species occurrence and abundance, Spatial distribution and Health of Ecosystems and Ecosystem Services, and a large amounts of Genomic Data generated through Next-Generation sequencing.
The main data types which will be generated are:

- Nucleic acid sequence data (including e.g. genomic breeding values);
- Taxonomic information and associated metadata;
- Molecular biological and field collection methods;
- Ecological information based on the results of experimental work;
- Camera trapping and GPS tag data;
- Vegetation data;
- Land use and land cover data (including socio-ecological information);
- Conventional biodiversity survey data (species presence/absence and abundance);
- Data regarding analysis methods (including computer modelling and simulation);

Each data type will be archived and disseminated in the ways that are appropriate to its specific qualities. Links for the different archives holding the main data types above will be available in a centralized location - the BioStudies database (https://www.ebi.ac.uk/biostudies/) - organized by project/study. Other data types will be archived directly in the BioStudies database, as well as in other more data specific databases and platforms, as outlined under the defined data types in the section below. Additional data types are likely to be introduced during the course of the BIOPOLIS project. The information

and details regarding this will be thoroughly monitored by the Data Management and Open Access Committee, and carefully outlined in the upcoming versions of the Data Management Plan (see further below).

3. DATA MANAGEMENT - HANDLING, STORAGE AND ARCHIVING

3.1. Nucleic acid sequence data

Nucleic acid sequence data can be archived in many ways depending upon the degree of analysis and annotation of features that has been undertaken on it. Raw reads and sample metadata will be archived in the European Nucleotide Archive (https://www.ebi.ac.uk/ena) which is part of the ELIXIR Core Data Resources (https://www.elixir-europe.org/). Sequences that can be associated with a clearly defined taxon and verified by a recognized taxonomic expert will be, whenever possible, also deposited in the GenBank (https://www.ncbi.nlm.nih.gov/genbank/) and the Barcoding of Life Database (http://www.barcodeoflife.org/). Protocols will be implemented so that submission to the above mentioned archives is performed consistently and, whenever possible, semi-automatically. The selected archives are reassuring that the data are easily found, openly accessible, reproducible, meaningful and can be reanalyzed by others.

3.2. Taxonomic data

Taxonomic information relating to sequences that are generated in this project will generally be archived in custom databases with links to the sequences and organisms associated with them. The custom databases will be hosted by the BIOPOLIS project website (https://www.biopolis.pt/) . In addition, the data on species occurrences will be added to the database of the Global Biodiversity Information Facility (http://www.gbif.org). Associated metadata, such as locations for type specimens or photographs, will be archived according to the requirements of the data repositories. As mentioned above, sequences that can be associated with a clearly defined taxon will also be deposited in BOLD or GenBank. Data generated within the scope of the project regarding population estimates etc will also be shared via e.g. the Living Planet platform (http://www.livingplanetindex.org/data_portal),

which covers not only the detection of a given taxon (as GBIF), but also allows to share actual population estimates and methods of assessment.

3.3. Molecular biological or field methods

Molecular biology or field collection procedures for biological materials will be archived either through peer-reviewed publication in Open Access journals, or by making a transcript or video of the methods and making this available on project page in the BioStudies database (https://www.ebi.ac.uk/biostudies/). This follows the H2020 principle of Open Research Data (ORD) publication.

3.4. Ecological results dissemination

Ecological information that is generated through the various projects that may be funded and run under the capacities of BIOPOLIS will be archived in a specific project page in the BioStudies database (https://www.ebi.ac.uk/biostudies/). This follows the H2020 principle of Open Research Data (ORD) publication.

3.5. Camera trapping and GPS-tagged animal data

The data/internally generated databases produced via camera trapping will be stored in available platforms such as eMammal (https://emammal.si.edu), Agouti (https://www.agouti.eu) or, more recently, Wildlife Insights initiative (https://www.wildlifeinsights.org). The advantages and limitations of each platform will be assessed to select the most appropriate repository. If considered adequate to ensure optimal dissemination of the data, more than one platform may be utilized. Animal movement data generated via GPS-tagging will typically be archived and managed through the movebank platform (https://www.movebank.org/cms/movebank-main). These types of data, in particular camera trapping which consist of images, will require large amounts of memory for storage and processing. Therefore, *in situ* storage will be used, in combination with an integrated management process through one or several of the above mentioned archive platforms.

3.6. Vegetation data

In-situ vegetation plot data such as vegetation cover, species abundance, height, DBH, phenological stage etc will be archived in a number of databases, e.g. https://www.givd.info/, http://www.sivim.info/sivi/and https://www.gbif.org/. The possibility of establishing permanent forest vegetation plots will also be explored and data made available through https://www.forestplots.net/pt. Remotely sensed vegetation data such as spectral vegetation indices, biophysical parameters, vegetation cover, productivity, phenology, seasonality and spatial heterogeneity will be archived in for example https://zenodo.org/, https://github.com/, https://datadryad.org/ and https://www.re3data.org. As for the latter data type, advantages and limitations of each platform will be assessed, and the best option(s) selected to comply with the FAIR principles.

3.7. Land use and land cover data

There are ongoing efforts by the Food and Agricultural Organization of the United Nations (FAO; fao.org) regarding Land Use and Land Cover (LULC) data, and the organization and harmonization of this data in the databases. Examples here are the Global Land Cover-SHARE (GLC-SHARE) database (http://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1036355/), Global Land Cover 2000 (GLC 2000; Global Land Cover 2000 database. European Commission, Joint Research Centre, 2003; https://forobs.jrc.ec.europa.eu/products/glc2000/glc2000.php). LULC data generated within the scope of BIOPOLIS gathered via e.g. aerial photographs, GPS and Geographic Information System (GIS) and Remote sensing (RS) may be archived in databases such as the example indicated above. An assessment of the suitability of different platforms and archives will be performed.

3.8. Conventional biodiversity survey data

Data gathered regarding e.g. species presence/absence, species abundance and other data in this context will be handled using the Darwin Core Standard (DwC), which offers a rigid, simple and flexible structure for compiling biodiversity data from a varied and variable range of

sources. This archived data will then be released and shared, following the publishing of the corresponding paper, via e.g. https://www.gbif.org/.

3.9. Data analysis methods

Computational data analysis methods, such as scripts for data processing, statistical analysis and graphical display of data will be archived either as supplementary material for published papers in Open Access journals, or, when possible, by placing them in the BIOPOLIS github page (https://github.com/biopolis). See SOFTWARE section below for further details.

In parallel with the above outlined processes in handling and storing the gathered project data, and to further increase the capacity to handle, process and store big data, particularly the data generated through the omics platforms, an upgrade of the Storage Area Network, the Compute Nodes, and an upgrade of the connection to the National computation platform (at University of Minho) will be executed via acquisition of additional equipment to establish a robust and well-functioning BIOPOLIS Computational Platform, as detailed in the project Grant Agreement.

4. DATA DISSEMINATION – SHARING AND PUBLISHING

The BIOPOLIS project will follow the H2020 FAIR principles to make its data findable, accessible, interoperable and reusable. The data (raw and processed) for each project that is conducted under the BIOPOLIS umbrella will be accessible through the BioStudies database (https://www.ebi.ac.uk/biostudies/). This database can hold descriptions of biological studies, links to data from these studies in other databases (e.g., GenBank, ENA, BOLD), as well as data that do not fit in the structured archives (e.g. specific tables with results, protocol descriptions, photos, videos). Furthermore, the database also enables the submission of supplementary information and links to it from companion publications. The BIOPOLIS project website will further contribute to data dissemination by providing for each project the corresponding page in BioStudies database with the links for all data and metadata. Additional sample/data specific databases will also be used in the sharing and publishing processes, such as the databases mentioned in the section above regarding camera trapping data, GPS-tagged animal data, and vegetation plot and remotely sensed data.

Data collected or produced under BIOPOLIS will be made available publicly for reuse. To this end the data will be shared with one of the Creative Commons Licenses. By default, the data will be licensed using Creative Commons (CC) license Attribution 4.0 International license (https://creativecommons.org/licenses/by/4.0/) which allows third party researchers the right to redistribute the material in any medium or format and build upon the material for any purpose, even commercially. However appropriate credit must be given to the original source, provide a link to the license, and indicate if changes were made.

The BIOPOLIS project will, as outlined above and in the GA, produce a wide range of exploitable results and data, which will contribute to the generation of new research, but also in addressing societal challenges and in developing novel products, processes and services. To promote and facilitate the dissemination and exploitation activities regarding the key results and generated data, these activities will be managed and monitored by the Communication, Advancement and Engagement Unit, which is part of the organizational structure of BIOPOLIS. Within this unit at least two officers will be dedicated to results and data communication with proven expertise in Knowledge Transfer and Business Relations. These officers will also liaise with researchers and other officers (including the Communication and Dissemination Officers that will be hired within the scope of BIOPOLIS at the partner University of Montpellier), to effectively promote the communication and exploitation of obtained results and data on both national and international levels. These activities will all be concerted in the overall scope of the well-defined dissemination and exploitation strategy, outlined in the project Communication, Dissemination and Exploitation Plan (D7.2; First version due at M12).

5. SOFTWARE

All software or routines developed during the project will be available in GitHub or GitLab. The BIOPOLIS project will follow well established standards for packaging, versioning, documentation, updating and installation. The source code of the software will be freely available in code repositories under permissive open source licenses (by default GNUGPLv3; Apache License, Version 2.0, or MIT). These routines will be continuously updated by the BIOPOLIS IT Experts/IT Committee as the project progresses, and these updates will be carefully detailed in the following versions of the project DMP.

6. DMP IMPLEMENTATION AND MANAGEMENT OF DATA

As detailed in the previous sections of this DMP, all data and analysis methods will be archived in well-established public databases thus ensuring long term preservation of the data. Many of the selected archival databases are certified repositories and are part of the ELIXIR core infrastructure. Although there is no cost to store data in these databases, there will still be a cost for making data FAIR - the time spent by researchers and/or technicians to upload and curate the data. The goal is to minimize this time, and therefore the cost, by defining clear internal semi-automatic procedures to make the data FAIR.

The BIOPOLIS Data Management and Open Access Committee, selected by the Science Council and the Coordinating Researcher of each Research Unit and approved by the Board of Directors, will be responsible for the overall development of the DMPs, and the planning, implementation and monitoring of all aspects regarding data management and open access policies and procedures attached to it, as well as the review and updating of the DMPs according to timelines detailed the in the project GA . The committee will consist of five members, and include Senior Researchers as well as IT Experts (linked to the BIOPOLIS IT Committee), with relevant experience in both the management of data and Open Access matters.

The DMP will be widely shared and advertised among all BIOPOLIS participating researchers/partners, and the committee will advise on technological solutions regarding data preservation and sharing in the scope of the overall project.

The procedure for checking that the DMP is followed will initially encompass the following steps: i) a spreadsheet (in a predefined format) with a description of all data associated to the project activities as well as for manuscripts regarding publications will be provided to the Data Management and Open Access Committee; ii) the Data Management and Open Access Committee will validate the information provided and check that it is compliant with BIOPOLIS's DMP and give the appropriate feedback; iii) once the compliance with the DMP is confirmed, the Data Management and Open Access Committee will make sure that all required links to the project's page in the BioStudies database, as well as to any additionally used databases, are added to the official BIOPOLIS website. A compliance check will be made before submitting the companion manuscript. Nevertheless, the data may only be made publicly available after the manuscript has been accepted and published. The DMP compliance

procedure will continuously be developed and updated by the Data Management and Open Access Committee, and this will be carefully outlined in the next versions of the BIOPOLIS DMP.

When required, specific training on FAIR data and utilization of the selected archives will be provided by organizing workshops covering relevant topics.

Costs related to data management, curation and preservation will generally be covered via research and innovation project budgets directly associated with BIOPOLIS, whereas other additional costs related to this will be covered by the BIOPOLIS annual budget.

7. DATA SECURITY, RESAERCH ETHICS AND ANIMAL WELFARE

Data security is ensured by archiving the data in established public repositories, as described in this DMP, and will be monitored by the Data Management and Open Access committee.

A consortium agreement will be put in place between the project partners to manage the ownership and access to key knowledge, e.g. in processes of pursuing market/business opportunities which materialize through project activities. In connection to this and in the scope of the BIOPOLIS Knowledge Management System - KMS (D7.4a), a multilevel access permission system will be installed and implemented to reduce the possibility in wrong uses of the data, as well as protecting the interest of the partners.

Research Ethics considerations and full compliance with the Nagoya Protocol on Access and Benefit Sharing (ABS), EU Regulation No 511/2014 and Portuguese Cecree-Law 1222/2017, related to the collection and analysis of genetic samples and the generation of the BIOPOLIS project data is all thoroughly outlined in the BIOPOLIS Grant Agreement in the Technical Annex (under Ethics and Security). This is the responsibility of the BIOPOLIS Ethical and Animal Welfare Committee, which will guarantee that all the research is conducted following the best practice and the highest international standards. As outlined in the GA, this committee, will together with the "Centre d'Elevage etv de Conditionnment Experimental des Modè les Animaux" at University of Montpellier, develop the Guidelines in this area.

Moreover, and in the overlapping scope of the EU funded ongoing ERA Chair project – TropiBio and related to Third Countries , contacts have already been taken and a first meeting organized with Namibia, one of the countries in Africa where CIBIO already has established a TwinLab. Discussions were held and initial procedure planning made with representatives from the Ministry Turism and Environment, as well as from the Faculty of Agriculture and Natural Resources (UNAM, regarding Ethics Clearance/Permits and ABS issues. These procedures, or similar ones, will then be applied in additional countries in Africa, Europe and worldwide within the scope of the BIOPOLIS project, regarding these matters. Similar or related procedures regarding Data security, Research Ethics and Animal Welfare will also apply in the interactions and collaborations with other stake holders, such as larger private companies, SMEs and non-governmental organizations operating in the African countries, as well as in other parts of the world.

8. CONCLUDING REMARKS

In the scope of this Data Management plan, BIOPOLIS will be aiming to go beyond the standards of "best practice" for the research areas of Environment and Biodiversity assessment and Monitoring, Ecosystem Function and Restoration and Agrobiodiversity, Conservation and Competitiveness of local Genetic Resources and Farming systems. This will be achieved in a way that conforms to the H2020 programme's principles of FAIR data dissemination and that also follows an adequate ORD approach. Furthermore, the research fields encompassed under the BIOPOLIS umbrella are rapidly changing and, although carefully outlined in this DMP, the exact data types that will be generated are not yet fully determined when this first version of the DMP is due, and may therefore change/be expanded. The BIOPOLIS Data Management and Open Access committee will carefully follow changes in the research fields and adjust the data management plan when required and as necessary. This will be important if data archiving practices change within the scope of the wider research community. By exploring the public databases described in this Data Management Plan for the data archiving, and making sure that the specific links to the data and metadata are properly managed as described under Data Dissemination in this document , we will comply with the FAIR principles while also ensuring the availability of the data beyond the end of the BIOPOLIS project. In addition, the flexibility of many of the archiving databases used, will also minimize the adjustments in the data management when new types of data will be introduced and curated.

9. RELEVANT TECHNICAL REPORTS AND PUBLICATIONS

Below is a compilation regarding relevant Technical Reports and scientific Journal Publications in the context of this Data Management Plan for the project, taking advantage of the strong synergies between the BIOPOLIS project and currently ongoing EU funded ERA Chair projects EnvMetaGen and TROPIBIO, and the respective DMPs attached to those projects. Lists of project-specific relevant publications for the overall BIOPOLIS project were already outlined in the Grant Agreement (GA) and Description of Action (DoA) for the project related to CIBIO, as well as for the Teaming partners University of Montpellier UM), France and the Porto Business School (PBS), Portugal.

Technical Reports

EnvMetaGen Deliverable 4.2 (D4.2) Protocol for building and organizing reference collections of DNA Sequences. DOI: 10.5281/zendo.2586893

EnvMetaGen Deliverable 4.3 (D4.3) Protocol for field collection and preservation pf eDNA samples. DOI: 10.5281/zendo.2579806

EnvMetaGen Deliverable 4.4 (D4.2) Protocol for next-gen analysis of eDNA samples. DOI: 10.5281/zendo.2586885

EnvMetaGen Deliverable 4.5 (D4.5) Protocol for the processing of DNA sequencing data generated by next-gen platform. DOI: 10.5281/zendo.2586889

Journal Publications

Gonçalves J., Henriques R., Alves P., Sousa-Silva R., Monteiro A.T., Lomba Â., Marcos, B. & Honrado J. (2015) Evaluating an unmanned aerial vehicle-based approach for assessing habitat extent and condition in fine-scale early successional mountain mosaics. Applied Vegetation Science. doi: https://doi.org/10.1111/avsc.12204.

Ferreras, P., Diaz-Ruiz, F. & Monterroso, P. (2018) Improving mesocarnivore detectability with lures in camera trapping studies. Wildlife Research. doi: 10.1071/WR18037.

Ferreras, P., Díaz-Ruiz, F., Alves, P. C., & Monterroso, P. (2017). Optimizing camera-trapping protocols for characterizing mesocarnivore communities in south-western Europe. Journal of Zoology. doi:10.1111/jzo.12386.

Monteiro A.T.; Gonçalves J., Fernandes R.F., Alves S., Marcos B., Lucas R., Teodoro A.C. & Honrado J.P. (2017) Estimating Invasion Success by Non-Native Trees in a National Park Combining WorldView-2 Very High Resolution Satellite Data and Species Distribution Models. Diversity. doi: https://doi.org/10.3390/d9010006.

Monterroso, P., Alves, P. C., & Ferreras, P. (2011). Evaluation of attractants for non-invasive studies of Iberian carnivore communities. Wildlife Research. doi:10.1071/wr11060.

Monterroso, P., Alves, P. C., & Ferreras, P. (2014). Plasticity in circadian activity patterns of mesocarnivores in Southwestern Europe: implications for species coexistence. Behavioral Ecology and Sociobiology. doi:10.1007/s00265-014-1748-1.

Monterroso, P., Alves, P. C., Ferreras, P., & Fusani, L. (2013). Catch me if you can: diel activity patterns of mammalian prey and predators. Ethology, 119(12). doi:10.1111/eth.12156.

Monterroso, P.*, Brito, J. C., Ferreras, P., & Alves, P. C. (2009). Spatial ecology of the European wildcat in a Mediterranean ecosystem: dealing with small radio-tracking datasets in species conservation. Journal of Zoology. doi:10.1111/j.1469-7998.2009.00585.x.

Monterroso, P., Diaz-Ruiz, F., Lukacs, P., Alves, P.C. & Ferreras, P. (In press) Ecological traits and the spatial structure of competitive coexistence among carnivores. Ecology.

Monterroso, P., Godinho, R., Oliveira, T., Ferreras, P., Kelly, M.J., Morin, D., Waits, L., Alves, P.C. & Mills, L.S. (2019) Feeding ecological knowledge: The underutilized power of fecal DNA approaches for carnivore diet analysis. Mammal Review. doi: 10.1111/mam.12144.

Monterroso, P., Rebelo, P., Alves, P.C., & Ferreras, P. (2016). Niche partitioning at the edge of the range: a multidimensional analysis with sympatric martens. Journal of Mammalogy. doi:10.1093/jmammal/gyw016.

Monterroso P., Rocha, L.F., Van Wyk, S., António, T., Chicomo, M., Kosmas, S., Lages, F., Fabiano, E. & Godinho, R. (In press). Updated ranges of the Vulnerable cheetah and Endangered African wild dog in Angola. Oryx. doi: 10.1017/S0030605319000966

Nagendra H., Lucas R., Pradinho Honrado J., Jongman R.H.G., Tarantino C., Adamo M. & Mairota P. (2013) Remote sensing for conservation monitoring: Assessing protected areas, habitat extent, habitat condition, species diversity, and threats. Ecological Indicators. doi: https://doi.org/10.1016/j.ecolind.2012.09.014.

Ortega M., Guerra C., Honrado J.P., Metzger M.J., Bunce R.G.H.& R.H.G.Jongman R.H.G. (2013) Surveillance of habitats and plant diversity indicators across a regional gradient in the Iberian Peninsula. Ecological Indicators. doi: https://doi.org/10.1016/j.ecolind.2012.12.004.

Pôças I., Gonçalves J., Marcos B., Alonso J., Castro P. & João P. Honrado J.P. (2014) International Journal of Geographical Information Science. doi: https://doi.org/10.1080 /13658816.2014.924627.

Torres J., Gonçalves J., Marcos B. & Honrado J. (2018) Indicator-based assessment of post-fire recovery dynamics using satellite NDVI time-series. Ecological Indicators. https://doi.org/10.1016/j.ecolind.2018.02.008.

BIOPOLIS

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Grant Agreement Number 857251.