THE PORTUGUESE INFRASTRUCTURE FOR BIOLOGICAL DATA

ANNUAL REPORT

2020

BioData.pt

elixir PORTUGAL

CONTENTS

We are pleased to introduce this year's Annual Report, which illustrates the shared commitment of the teams and communities congregated under the Portuguese ELIXIR Node, Biodata.pt, to creating an excellent research infrastructure for biological data, supporting research, training and innovation.

In addition to a core set of services, we now have established National communities along with new ones that are emerging, which reflect our strategy of engaging with other communities, at European and International levels, equally committed to developing and adopting FAIR data principles for addressing the complexities of biological data management and coping with the ever increasing volumes of data.

As we approach the end of the initial funded projects for implementing the Biodata.pt, the National biological research infrastructure, and prepare to celebrate out 5th anniversary as ELIXIR Node, we now look back to 2020 as the year of development of the major strategic plans for evolving and adapting Biodata.pt to address the challenges of the next framework programme. In the pages that follow, you will see why our confidence is justified.

FOREWORD FROM OUR PRESIDENTS

José Pereira Leal and Mário Gaspar da Silva

OUR CEO'S STATEMENT

Ana Portugal Melo

The year 2020 will be forever and worldwide labelled by the COVID-19 pandemic that urged us to adapt. We had started the year by delivering several face-to-face capacity building and networking events, setting sails for what we planned to be a very intense and interactive year. And then… COVID-19.

Sent home by March, our being a virtual and distributed Research Infrastructure turned out to be an advantage, used as we were to meet and work remotely. We took time to better understand our users, designed and disseminated our portfolio of services, reviewed our website and adapted the materials of the “Ready for BioData Management?” programme to be delivered online. Meanwhile, the compute infrastructure continued to develop, a first prototype of the Local EGA was deployed, D-Cellerate, Ready for BioData Management?, Pegi3S and EvoPPI were recognized as new node services by the ELIXIR, a Knowledge Exchange Scheme was set up to transfer the MIAPPE standard to The Navigator Company, the CorkOak DB portal was launched, and…, and… The ELIXIR-CONVERGE project started and is working hard to deliver a toolkit of best practices in data management for life sciences. Noteworthy, the INCoDe.2030 certification was awarded to the “Ready for BioData Management?” programme.

A final word to share that we have been planning the future of BioData.pt in preparation for the next funding programme. We dream that in the next phase BioData.pt will be the Portuguese Platform of Biological Information organized into domain-specific Hubs to support Portuguese research & innovation organizations in data management and valorization for scientific, business and societal purposes, that will attract global visibility and access to Portuguese research.

Join us in this journey!

OVERVIEW

Recent years have seen a boost in data generation, particularly in the life and health sciences, that is supported by the progressive availability of technology (i.e., equipment, connectivity and analytic power). These large amounts of data—the so-called big data—have nowadays disrupted the boundaries of social and economic sciences and are bursting from human and animal health, plants, marine resources and microbiology research, among others. Transforming biological data into information, and information into knowledge for the public good is one of the greatest challenges at the onset of the 20's decade. Moreover, democratizing the access to equipment, connectivity and analytic power by research organizations is enhancing the scientific process to produce results faster.

Extending skills, knowledge and resources to all researchers conducting their work in Portuguese organizations is the purpose of the National Roadmap of Research Infrastructures, and of the Portuguese distributed infrastructure for biological data - BioData.pt. This country-level effort is embedded, at the European level, in the European Strategic Forum for Research Infrastructures (ESFRI). ELIXIR, the European distributed infrastructure for biological data, is the landmark of ESFRI that brings together life science resources from across Europe, including databases, software tools, training materials, cloud storage and supercomputers. BioData.pt is the national node of ELIXIR, contributing to this European effort with computing, services and training resources, and channelling ELIXIR's resources to the Portuguese research community.

Research data go through a cycle of stages from the experiment design to publication and reuse, and care must be taken across all stages to preserve the integrity of the data and capture the metadata required for its interpretation and reuse, in a process called data management. According to the ongoing ELIXIR-CONVERGE project, the research data lifecycle comprehends planning, collecting, processing, analysing, preserving, sharing, and reusing. Generally accepted best practices suggest that these data should be findable, accessible, interoperable and reusable (FAIR). A recent study by the European Commission emphasized that the lack of best practices in research data management is costing the taxpayer at least €10.6 bn per year.

Reinforcing the goal of enabling scientists from all domains to take full advantage of the digital age resources in their research and aiming at a leadership role for Europe, the European Open Science Cloud (EOSC) was recently created, as an environment for hosting and processing research data to support science within the European Union. It fosters open science becoming the "new normal" in Europe and worldwide, with the corresponding impact in terms of transparency, effectiveness and efficiency. In spite of the public good goals of Open Science—namely the transformation of science through digital tools and networks, to democratize research making it more open, global, collaborative, creative and closer to society—it has limits. The four major limits to open science identified by the EOSC are privacy, security, sovereignty, and property. These are also major challenges for BioData.pt, as this research infrastructure is committed to collaborate with other organizations, namely with human health organizations and the industry.

BioData.pt supports the national scientific system in its research strategies and programmes through management and advanced analysis of biological data. It promotes scientific research in the agro-food and forestry, sea and health sectors, leveraging value creation projects based on biological information, in partnership with business research and innovation. Trust, collaboration, excellence and innovation are in our DNA, steering and shaping our actions as we are committed to respect and protect our users' data, work in partnership, make available our team of internationally recognized experts, and support the translation of knowledge into new services and products.

This report showcases the BioData.pt activities that contribute to a healthy and valuable management of biological data generated by Portuguese research & innovation organizations.

HIGHLIGHTS

10

Genome-scale metabolic models

BioData.pt systems biology working group from the University of Minho developed genome-scale metabolic models for the **cork oak tree** (in collaboration with ITQB-NOVA), ***Candida albicans*** (in collaboration with iBB-IST-ID) and other **industrially relevant microorganisms**

9

Computing Applications

BioData.pt developed 9 computing applications for scientifically relevant subjects: **Community Yeastract, Data Management Portal, GenomeFastScreen, GFFTool, Marine Metagenomics Annotation Portal, MEWPy, ProTReND, Python Video Annotator + idTracker AI**, and **SeagrassDB**

324

Participants in courses and workshops

The Portuguese research community participated in more than 20 capacity building activities organized by BioData.pt

12

International publications

The publication of 12 peer-reviewed articles in relevant scientific journals consolidated BioData.pt's international recognition

7 PEOPLE TRAINED AT ELIXIR TRAIN-THE-TRAINER PROGRAMME

1 FULLY OPERATIONAL COMPUTE INFRASTRUCTURE
CCMAR, IGC, TÉCNICO ULISBOA

2 NEW PROJECTS
VITACOV & CONVERGE

1 KNOWLEDGE TRANSLATION TO INDUSTRY
THE NAVIGATOR COMPANY

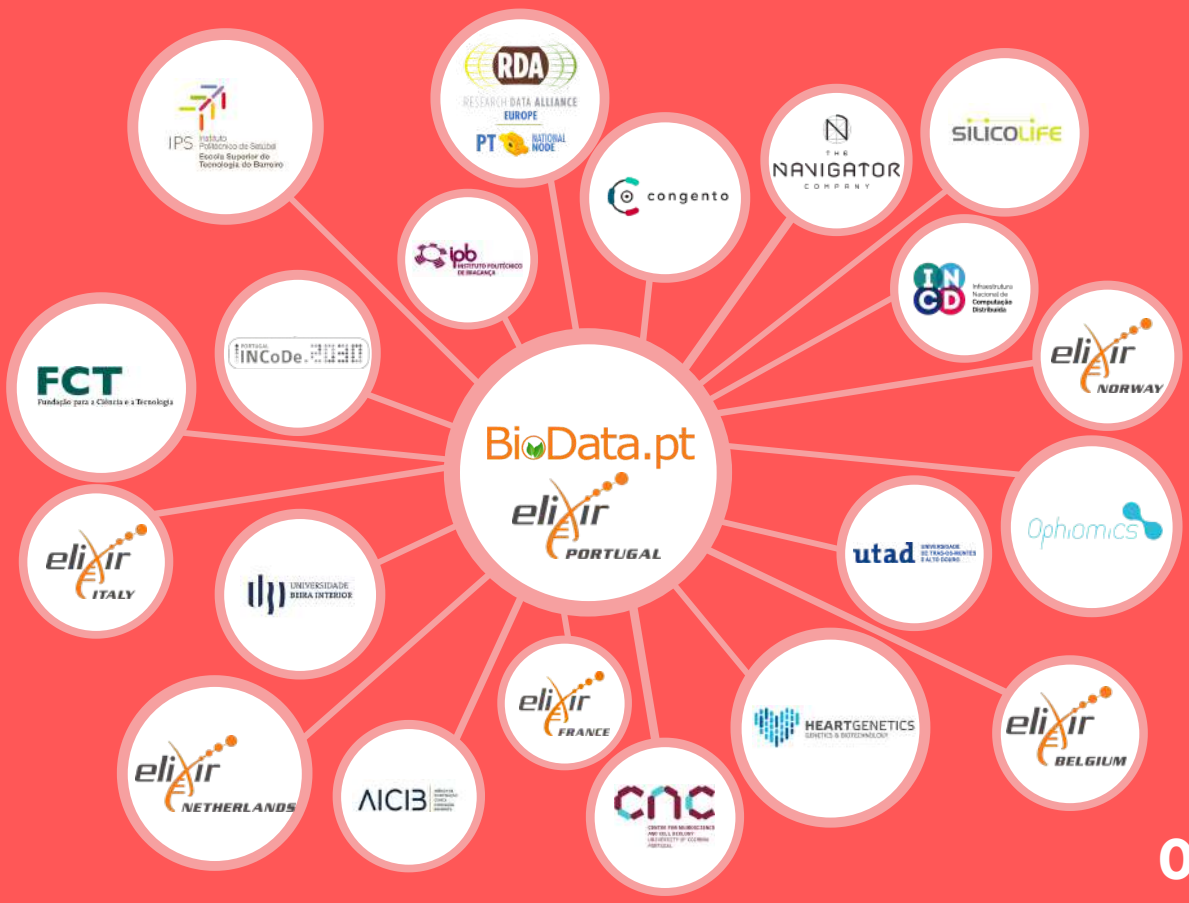
1 AWARD
INCODE.2030

22 ORAL COMMUNICATIONS

2 THESIS
1 PHD | 1 MSCI

4 POSTER COMMUNICATIONS

A SOLID AND BROAD NETWORK OF COLLABORATIONS



EMPOWERING USERS

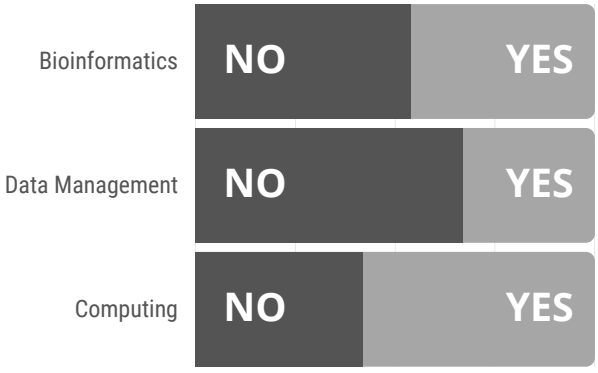
USER SURVEY

BioData.pt conducted a survey to identify the needs and skills of the Portuguese research & innovation community in computing, bioinformatics and data management.

› **GAP** between **NEEDS** and **SKILLS** on the topics of bioinformatics, data management and cloud computing!

› Groups **LACK SKILLS** on these topics and **LACK ACCESS** to dedicated services in their institutions!

DEDICATED SERVICES IN R&I INSTITUTIONS



RESEARCHERS' NEEDS

> 55% Training

> 40% User support and consulting services

~ 30% Computational resources

New ELIXIR Node Services

D-Cellerate

EVOPPI

Ready for BioData Management?
Capacity Building for the Life Sciences

Bioinformatics Docker Images Project - Pegi3S, a collection of Docker images for commonly used bioinformatics software, **D-Cellerate**, a web application that provides a graphical interface for a popular single-cell RNA-seq package for R, **EVOPPI**, a web application that allows the easy comparison of publicly available data from the main protein-protein interaction databases for distinct species, and **Ready for BioData Management?**, a capacity building programme in data management for the life sciences, have been promoted to ELIXIR Node Services.

BIODATA.PT SERVICE HUB

- Computing Services
- Bioinformatics Services
- Data Management Services

In 2020, and supported by the information derived from the *Needs and Skills Survey*, BioData.pt reshaped its service provision and organized it in the **BioData.pt Service Hub**. This facilitates communication with the user that can identify and request the needed services via a dedicated portal.

BioData.pt offers **computing**, **bioinformatics** and **data management** services & training, to support the Portuguese research community in adding value to biological data.

The **BioData.pt Forum** was created to stimulate information exchange in scientific and technological BioData.pt topics.
This users' community problem solving tool is open to every researcher seeking troubleshooting and interaction with colleagues.

17161
Page views

239
Users

BIODATA.PT HUB FOR EDUCATION AND TRAINING

As a follow up to the *Needs and Skills Survey*, the **BioData.pt Hub for Education and Training Information** was launched to concentrate the offer in education and training in computing, bioinformatics and data management in a **UNIQUE** access point.
The BioData.pt Hub for Education and Training Information is aimed to be a participatory hub, where **YOU** can contribute by sharing related courses and initiatives of **YOUR ORGANIZATION**!

BIODATA.PT COMMUNITIES

PLANT SCIENCES

The Plant Sciences endeavor to understand how plant phenotypes arise from the play between genotype and environment, which is critical to address the challenge of sustainable production of food and non-food plant products (e.g. wood, paper, cork) in the face of climate change, as well to manage natural ecosystems and fight climate change.

Formed at the onset of the ELIXIR-EXCELERATE project, the Plant Sciences Community aims at realizing the vision underpinning the FAIR data principles: enabling integration and reuse of plant phenotyping and genotyping datasets, so that full value can be extracted from the increasing wealth of data collected by researchers and producers worldwide.

In few domains is this realization so challenging, because few domains are so heterogeneous at so many levels. Plant experiments can range in setting from growth chambers and greenhouses to outdoor cultivated fields and wild forests, which differ fundamentally in the level of detail and manner with which environmental conditions need be recorded. Additionally, data can range in scope from the molecular to the organismal to the environmental level, and can be collected and processed fully manually, semi-automatically, or fully automatically.

During the ELIXIR-EXCELERATE years, the national Plant Sciences Community co-lead its European counterpart, having as main accomplishments the publication of the metadata standard MIAPPE version 1.1 and the release of the data exchange standard BrAPI version 2.0. Together, these resources address the key FAIR criteria where the domain was lacking: MIAPPE specifies the metadata fields that must be provided, and how they should be filled in, for a plant phenotyping dataset to be interpretable and reusable; and BrAPI specifies the means for accessing them.

Since then, the Plant Sciences Community has directed its efforts towards the adoption of these resources: by developing compliant data submission interfaces (ELIXIR Staff Exchange project); through training and outreach activities; by engaging with the industry (ELIXIR Knowledge Exchange project); and by incorporating these resources in recommendations and templates for data management plans (ELIXIR-CONVERGE). It remains one of the most active communities and a success story at both the national and European level.

BioData.pt's most active community, with success stories at both the national and the European level.

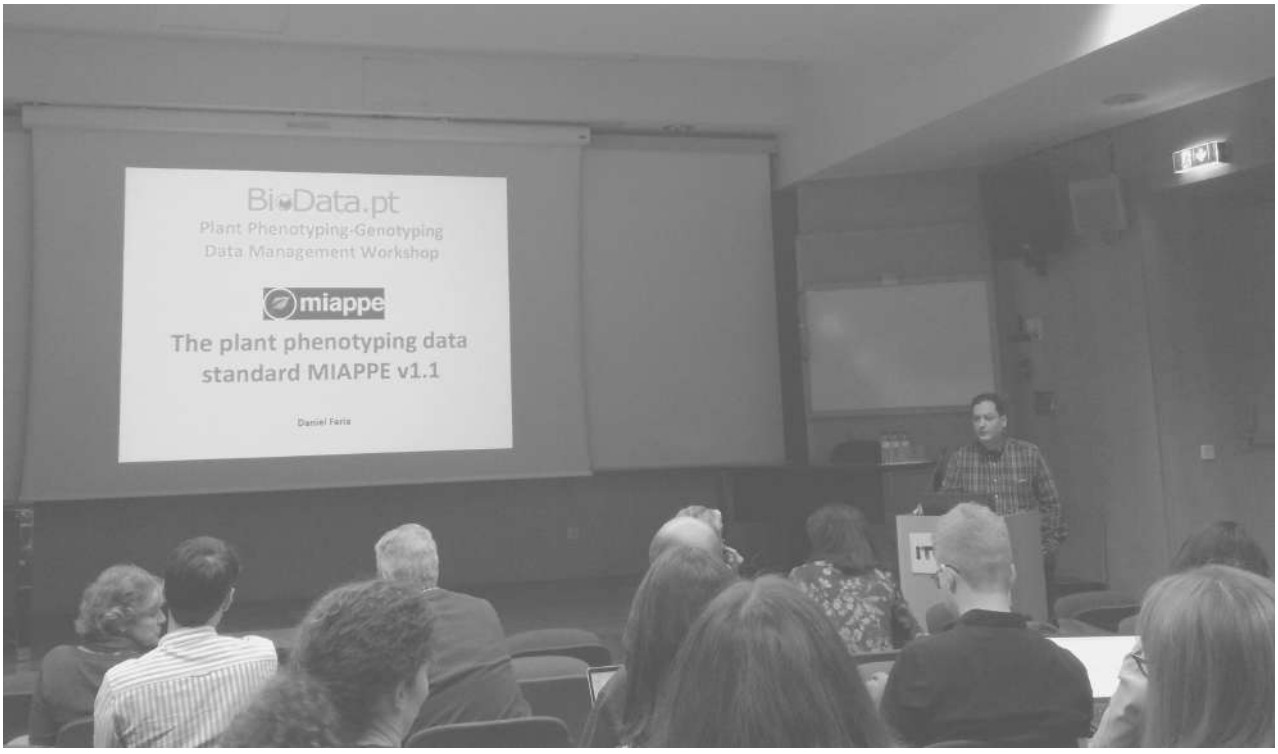
FAIR PLANT PHENOTYPING DATA SUBMISSION

The MIAPPE 1.1 release contemplated a solution for data submission based on ISA-Tab specification, which while functional, was not very user-friendly. On the side of BrAPI, calls for data upload are an integral part of the specification, but these are meant for computational use only—no interface is available for manual upload of datasets.

Given the aversion people tend to have to filling forms, and the complexity of MIAPPE in comparison with other "minimum information" standards, developing easy-to-use interfaces for MIAPPE-compliant data submission is imperative.

Led by BioData.pt, the ELIXIR Staff Exchange project **"Enabling FAIR plant phenotyping data submission through the Breeding API"** brought together the Belgian, Dutch, and French ELIXIR Nodes to tackle this issue.

This project had two aims. On the one hand, it aimed to incorporate MIAPPE into the popular data management platforms DataVerse and FAIRDOM-SEEK, which was already accomplished in 2020. On the other, it aimed to develop a stand-alone MIAPPE-compliant interface for data submission using BrAPI calls, which is in progress.



KNOWLEDGE TRANSFER TO THE PRIVATE SECTOR

A major hurdle to the global adoption of standards such as MIAPPE and BrAPI is demonstrating the value of such standards to the industrial sector. Many leading-edge companies in plant-related industries (e.g. wine, paper, wood, cork) include R&D branches that often dwarf academia in the volume of data they produce, but unless they comply with the same standards as academia, there is little hope for reconciling public and private data.

BioData.pt took a key step to overcome this hurdle in establishing a partnership with **The Navigator Company**, formalized in the ELIXIR Knowledge Exchange Scheme project: **"Bringing the ELIXIR Plant Data Infrastructure to the Portuguese pulp and paper industry"**.

This has led to the organization of The Navigator Company's plant phenotypic data according to the MIAPPE standard and to the submission of example datasets to the ELIXIR Portugal BrAPI end-point.

In this project, The Navigator Company agreed to share some of its datasets and publish them in BioData.pt's BrAPI end-point, whereas BioData.pt is undertaking the task of curating them according to the MIAPPE standard and converting them to linked data, which will facilitate the exploration of that data and its integration with other (public or private) datasets treated in the same manner.

This collaboration marks the **first time that the MIAPPE standard is adopted by the industry** and is also the **first knowledge translation to the industry led by BioData.pt**.

THE NAVIGATOR COMPANY

The Navigator Company is a leading force in the international pulp and paper market and one of Portugal's strongest brands on the world stage. In addition to its industrial activities, it carries out extensive research on Eucalyptus breeding and genetics, analysing approximately 300,000 data records, including genotypic and phenotypic data on over 60,000 individuals covering up to 4 generations of pedigree across a range of edaphoclimatic sites.

Another key partnership with the industry was established by BioData.pt in 2020. The **Phenospex** company—a leading provider of automated plant phenotyping systems with mainly academic clients—had the desire to ensure that the metadata captured by their information system is MIAPPE-compliant, as well as to implement BrAPI as a means to enable their users to access and retrieve data, as well as deposit it seamlessly to institutional BrAPI end-points. They looked towards ELIXIR experts, including members of the BioData.pt Plant community, to assist them with training and consulting.

Postponed by the COVID-19 pandemic, this collaboration is only starting in earnest in 2021, but it promises to be a major step towards the adoption of MIAPPE, as its inclusion in Phenospex's information system will spare the end-user of much of the burden of filling-in metadata.

CORKOAKDB

CorkOakDB aims to integrate the knowledge generated from fundamental and applied studies about Quercus suber, with a focus on genetics.

Cork oak is a crucial species to the Portuguese economy and identity, with ongoing efforts for its improvement, management and conservation. Studying the genetic structure of cork oak is essential for the success of these efforts, requiring the study of genes controlling traits of interest.

CorkOakDB aims to be a reference portal for scientific research on cork oak, aggregating all available genomic and transcriptomic data on this species. It offers a set of standard tools for data visualization and retrieval that enable core genomics analyses such as candidate gene identification for functional studies.

In 2020 the CorkOakDB developing team invested in the addition and integration of new transcriptomic datasets and curated functional information from specific genes, retrieved from published scientific research.

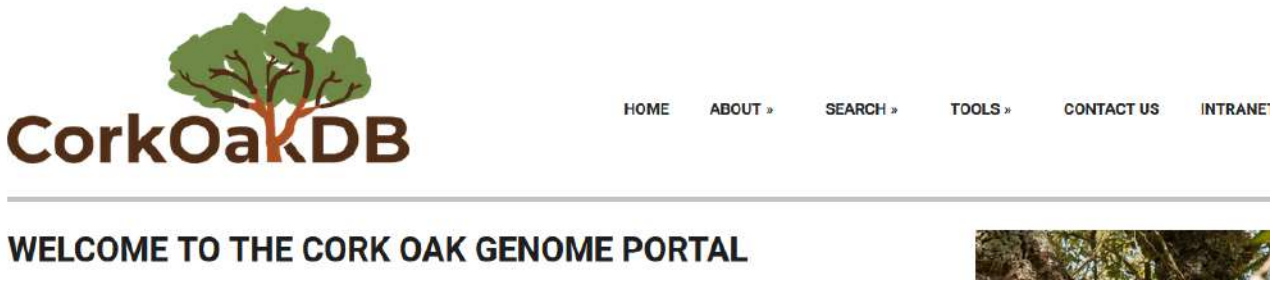
1452	7862	1879	1
Users	Page Views	Sessions	Publication

OFFICIAL LAUNCH!

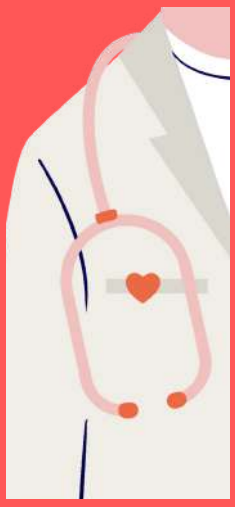
On the 30th of March of 2020, we oficially launched CorkOakDB, releasing a series of interviews and tutorials on our YouTube channel.

PUBLISHED!

Published on the last day of 2020 the article "CorkOakDB—The Cork Oak Genome Database Portal" represents a major step in getting the importance of this resource recognized.



HEALTH DATA

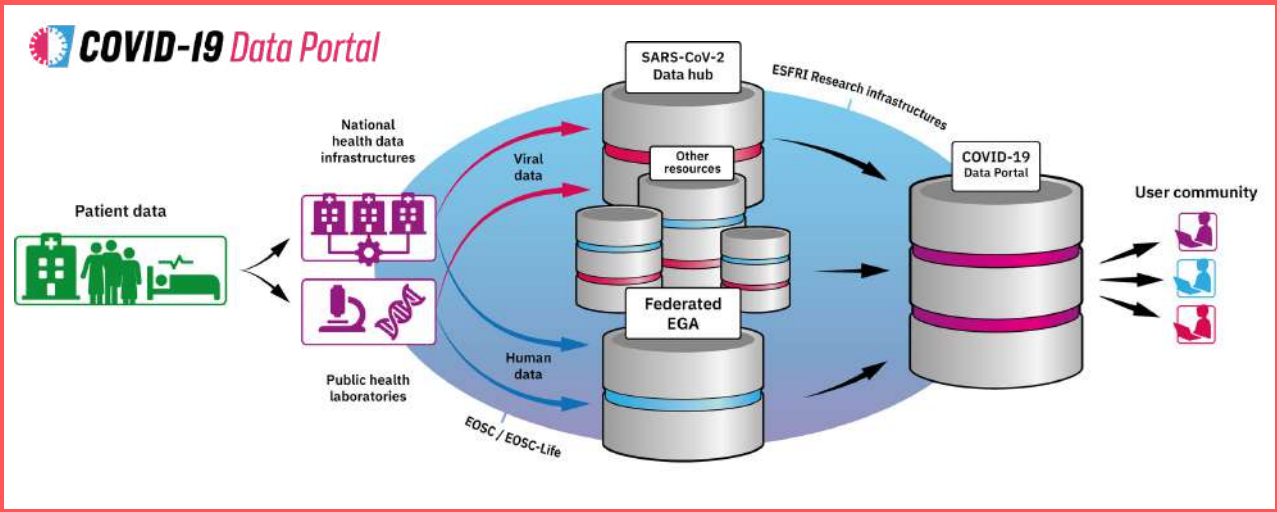


Solid steps for a healthy future...

According to the **Global Alliance for Genomics and Health** (GA4GH), the ongoing endeavor of -omics data sharing is becoming paramount in specific niches of clinical care allowing for a thorough diagnosis and the development of tailored treatment strategies (e.g. rare diseases and some cancer types). However, such benefits will only be available to the general population if researchers and clinicians can access and make comparisons across data from millions of individuals and that requires a strong collaboration between several stakeholders, such as research institutes, hospitals and other public health services. These challenges gained particular traction with the current COVID-19 pandemic.

IN EUROPE

At the European level, a COVID-19 dedicated portal concentrating SARS-CoV-2 information and tools was deployed in a joint effort conducted by ELIXIR, providing real-time information to advise decision-makers in managing the pandemic situation. National portals are being deployed to accelerate the contribution of all member states. The European **COVID-19 data portal** also aims to federate the genomes of COVID-19 patients. Such federated repositories may benefit from the **Federated European Genome-Phenome Archive** (FEGA) supported by the ethics protocol of GA4GH. The FEGA is a distributed network of repositories for sharing human -omics data and phenotypes. The deployment of national instances of FEGA will allow the discovery of biomedical datasets across Europe, overcoming ethical issues and the hurdles of national privacy laws and regulations, and their remote analysis, whilst retaining the data access control under the authorities of the country of origin.



IN PORTUGAL

At the national level, ELIXIR PT has deployed a prototype of the **local EGA** that could accept real genomes. This solution could federate this information so that each health data provider unit maintains its own repository, connected to the Portuguese National EGA, which in turn connects all national information to the EGA. Critical ethics issues are yet preventing the use of the EGA, and a solution at the European level is being sought. Further to ethics are the physical, technical, human resources and financial requirements that each health provider needs to maintain such a system.

In addition to the prototype of the local EGA, BioData.pt's contribution to the national health data ecosystem is now extended in a **service of electronic case report forms** (**eCRF**) **repository** to facilitate clinical research projects, such as **VITACOV**, and a **dedicated human data hub**, in the **Data Management Portal**.

These health data management tools can make data available for further research in human health.

VITACOV PROJECT

VITACOV: Vitamin D-related polymorphisms and vitamin D levels as risk biomarkers of COVID-19 infection severity was a project led by Hospital de Santa Maria in collaboration with HeartGenetics and other health and research organizations where BioData.pt was asked to contribute with the provision of an eCRF.

It aimed at understanding if an association exists between polymorphisms in vitamin D-related genes, vitamin D levels and the disease severity.

MARINE RESOURCES

The Marine Resources Community provides support, services and a forum for interaction among researchers and data analysts in the marine domain. These activities generally pivot about the analyses of long-term observational and ecological data, and the manipulation of genomic and genetic sequence data.

> **“The Marine Resources Community provides a focal point for marine biologists to collaborate and promote open science and FAIR data standards”**
>
> **Cymon Cox**
> **(Marine Resources Community Coordinator)**

SEAGRASSDB

Seagrasses are a group of approximately 86 species of marine flowering plants providing important ecosystem services. Uniquely among flowering plants, adaptation to the marine environment imposes major constraints on their morphology, structure and physiology. Importantly, seagrass traits are early indicators of environmental quality and change, with changes in these traits occurring before population level responses, species composition and associated biodiversity. Consequently, seagrass habitats are prime indicators of ecosystem health and quality, while functional plant traits can be used to assess ecosystem processes and services.

The **SeagrassDB** is a digital platform that allows researchers to share phenotypic trait data from seagrasses. A total of 206 traits (e.g. biochemistry, growth and development, morphology, physiology) are registered, and the metadata describing an observation is compliant with the MIAPPE standard. SeagrassDB is built upon the principles of Open Science and the promotion of FAIR data standards. The website is open to registration by everyone that wishes to contribute seagrasses phenotypic data. Public data sets are accessible without registration.

Seagrass DB

MICROBIOLOGY AND BIOTECHNOLOGY

The Microbiology and Biotechnology Community is focused on the development of software tools and web-based services for metabolic model reconstruction, phenotype simulation and strain optimization using constraint-based models, as well as their integration with omics databases. As a result, a few services have been made available.

SSDB & PHAGEPROMOTER

The Strain Design Database (**SSDB**) is a collection of genetic modifications and *in silico* computational strain optimizations. **PhagePromoter** is a tool developed for locating promoters in phage genomes.

BIOISO & TRANSYT

The Biological networks constraint-based In Silico Optimization (**BioISO**) is a tool aimed at assessing the connectivity of metabolic networks. The Transport Systems Tracker (**TranSyT**) is a solution aimed at identifying genome-wide transport systems.

MERLIN

The metabolic models reconstruction using genome-scale information (**merlin**) is a graphical and user-oriented software framework for reconstructing genome-scale metabolic models. In 2020, a model was published and over 10 are under on-going internal reconstruction. Both TranSyT and BioISO are available as merlin plugins.

MEWPY

Metabolic Engineering Workbench in Python (**MEWpy**), is a computational strain optimization tool that can search for the best strategy to maximize the production of a target compound, searching for different types of genetic modifications and being able to work with different types of metabolic models and simulation approaches.

THE YEASTRACT+ PORTAL

The **YEASTRACT+** portal has been continually updated for more than 15 years, currently containing 10 yeast species grouped into 3 distinct, but interconnected databases: Yeastract; PathoYeastract; and N.C.Yeastract. YEASTRACT+ provides bioinformatics tools for the prediction and visualization of gene and genomic regulation based on orthologous regulatory associations described for yeast species, based on comparative genomics.
Recently, the community created Community YEASTRACT, a tool for the analysis of genome sequences. Here, from a genome assembly of interest at NCBI, and with a minimum set of commands, the new genome is compared with a given genome of reference in search of homologous genes, shared regulatory elements and predicted transcription associations, thus benefiting from immediate access to all the comparative genomics queries offered in the YEASTRACT+ portal.

BIODATA.PT PLATFORMS

BioData.pt Platforms coordinate the provision of services to support the management of data generated by the research activities of BioData.pt communities. These are mirrors of the ELIXIR's of which BioData.pt has prioritized **Compute**, **Interoperability** and **Training**. The work developed in the scope of the Interoperability Platform is covered under the Plant Sciences and Data Management sections. Here, the Compute and Training Platforms will be addressed.

COMPUTE

The **BioData.pt Computing Service** is an ELIXIR node service that provides a full stack of computational resources to the Portuguese life science research community, complemented by a dedicated user support. The BioData.pt computational infrastructure, implemented over a hybrid cloud and high performance computing resources, supports access, usage and storage of digital objects and is coordinated by INESC-ID. This infrastructure includes **CCMAR**, **IGC** and **Técnico ULisboa** as cloud service providers, and constitutes the backbone behind most of BioData.pt's activities. Additionally, as a member of the Galaxy Pulsar Network, the BioData.pt computational infrastructure provides computing capacity to Galaxy Europe, thus enabling Portuguese researchers to use the Pulsar Network resources.

CCMAR Centro de Ciências do Mar	INSTITUTO GULBENKIAN DE CIÊNCIA	TÉCNICO LISBOA
545 vCPUs	**400** vCPUs	**1567** vCPUs
1.8 RAM (TB)	**5.1** RAM (TB)	**8.7** RAM (TB)
18 Disk Storage (TB)	**112.6** Disk Storage (TB)	**130.7** Disk Storage (TB)
25 Active Instances	**19** Active Instances	**33** Active Instances

TRAINING

The provision of training in BioData.pt capitalises on more than 20 years of face-to-face hands-on courses. In the beginning of 2020, BioData.pt has raised the training room at IGC to a premium quality level. It has kept the ability to support a wide variety of training courses, now enhanced by better networking, audio and video capture and playback capabilities, and a new area especially suited for collaboration working in small groups. This environment allows easy replication and remote training delivery using distance learning techniques for the Portuguese research community.

In 2020, materials for 21 training courses, designed for intensive face-to-face delivery with hands-on practice, have been assembled in a **publicly available collection**, composed of both a repository and a website to explore the content. Replication of these courses in an appropriate training environment requires little more than engaging instructors and minimal organizational effort, thus unlocking the potential for delivering training in a delocalized way, while benefiting from the design of the face-to-face training programme. BioData.pt has innovated in assembling a collection of course materials with a common design method, and making it publicly available under a CC-BY license.

The provision of training in BioData.pt capitalises on more than 20 years of face-to-face hands-on courses

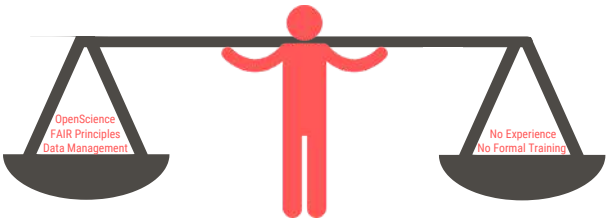
2020, the year of the COVID-19 pandemic first worldwide consequences, came with severe restrictions in offering face-to-face events of all kinds, obliging the organisers, instructors and learners to engage in serious efforts to adapt to online delivery, both in content and in format. BioData.pt has been preparing to offer many of its courses and workshops in both online and face-to-face versions. A serious effort is in place to design or redesign training for online audiences, rather than just carelessly playing unmodified courses in a video sharing platform. BioData.pt, as the Portuguese node of ELIXIR, was pivotal in the development and delivery of the first edition of the ELIXIR Train-the-Trainer for online audiences, hosted in October 2020. In the near future, BioData.pt will engage in developing training for e-learning on fixed and mobile platforms.

RESEARCH DATA MANAGEMENT

DATA MANAGEMENT FOR HEALTH AND LIFE SCIENCES

The FAIR data principles are being enforced by funding agencies and publishers as they represent the solution to the knowledge discovery and data reuse problem in the age of big data. Indeed, an analysis by the European Commission estimates that adherence to the FAIR principles could save the European economy at least €10.2bn every year. The problem is that complying with these principles requires good data management practices throughout the lifecycle of the data, a burden which currently falls mainly on the shoulders of the researcher, seldom well versed in data management.

BioData.pt recognized the need to bring awareness about data management, and capacitate and support the Portuguese scientific community (in both academia and industry) in carrying out data management activities.



Scientific Community

ELIXIR-CONVERGE

Funded by the European Commission, the **ELIXIR-CONVERGE** is a project fully devoted to data management, which aims to standardise and support life science data management across Europe.

This European-wide effort brings together experts from all 23 ELIXIR Nodes to work collaboratively on some of the major challenges in data access, provisioning and distribution.

BioData.pt, as the Portuguese node, is deeply involved in this project, as an active member of the data management expert network (WP1), by engaging in training and capacity building activities on this topic (WP2), by co-developing and serving in the editorial board of a data management toolkit to support researchers (WP3), by collaborating in communication, impact, and industry activities (WP4), by leading the plant science demonstrator selected for the project (WP5), by participating in project management (WP6), and more recently, by setting-up data management infrastructures to support COVID-19 research (WP7).

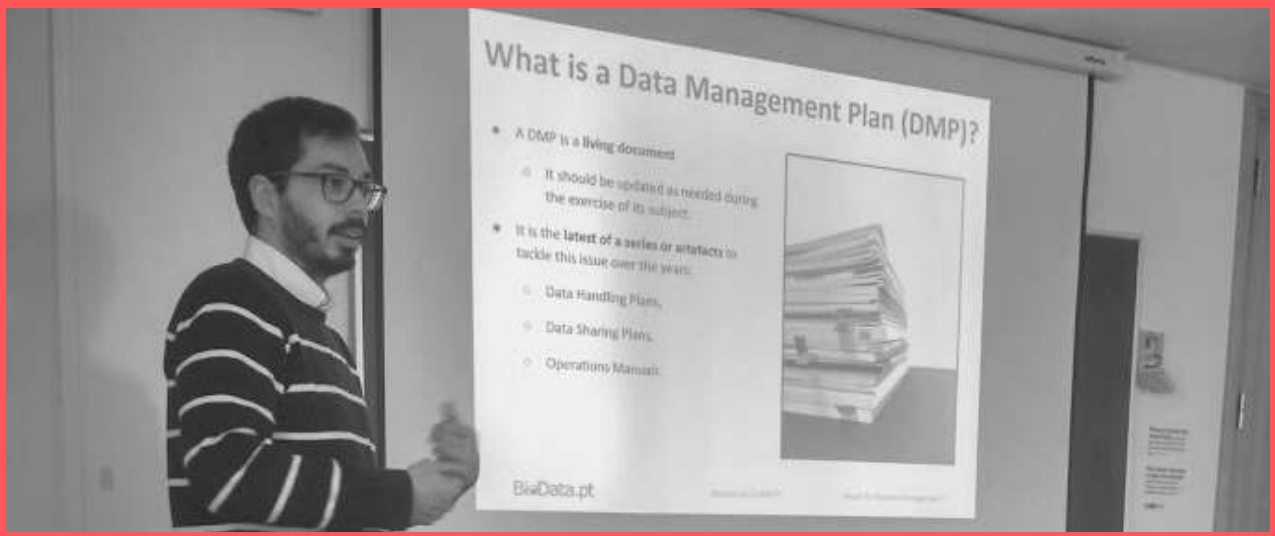
> **“Many ELIXIR Nodes already provide local support and expertise with data management to national research projects and life scientists in that country. With the ELIXIR-CONVERGE project, we want to ensure that this data management provision is as developed as possible within each country and as connected as it can be across countries too.”**
>
> **Niklas Blomberg (CEO at ELIXIR)**

DATA MANAGEMENT PORTAL

Launched in 2020, BioData.pt's **Data Management Portal** is an instance of DataVerse, a digital data management platform for data deposition, sharing, and publication. While the portal can, in principle support all research domains, BioData.pt aims to specifically support COVID-19 research, other Human Health research, and the Plant Sciences.
Researchers or organizations can use the Data Management Portal to deposit, annotate, share and publish their data, as well as to collect public data for reuse.
The portal encompasses a number of metadata standards which researchers can choose from, as appropriate for their domain, including the MIAPPE standard, for which BioData.pt developed a DataVerse template in 2020. The BioData.pt Data Management Portal serves therefore as a vehicle to support data management activities and foster the adoption of the FAIR data principles.

READY FOR BIODATA MANAGEMENT?

Launched in 2019, BioData.pt's capacity building programme in data management for the life sciences, "Ready for BioData Management?", really took off in 2020, with 12 training events realized. Moreover, 2020 saw the programme be endorsed and sponsored by the Portuguese node of the Research Data Alliance, as well as be endorsed by the INCoDe.2030 initiative.
The programme aims to empower researchers and institutions to manage their data more effectively and efficiently, as well as comply with funders' and publishers' demands for FAIR and open data.
The training offer is fully modular, with introductory and advanced material on data management planning which can be offered in the form of only theoretical lectures, independent 1-day hands-on courses, or an integrated 2-day event. A course on day-to-day data management is currently being planned to complement them. Due to the COVID-19 pandemic, all the materials have been adapted to be delivered fully online.



PERFORMANCE & IMPACT

PERFORMANCE & IMPACT ASSESSMENT OF BIODATA.PT ACTIVITIES

Demonstrating performance and impact of publicly funded projects is central to showing respect by citizens and stakeholders, as well as ensuring credibility and long-term sustainability of research infrastructure activities.

“In 2020, the Portuguese Node of ELIXIR showed incredible dynamism and empowerment on the impact evaluation front. I was particularly impressed with the leadership shown by this national Node in the context of an ELIXIR-funded Staff Exchange project, and it was an absolute pleasure to take part in the two impact workshops they hosted in 2020. In 2021, my hope is that ELIXIR Portugal will continue in this direction, so that the skills and experience they have acquired benefit other ELIXIR Nodes, in support of their long-term sustainability.”

Corinne Martin (External Relations Officer at ELIXIR)

AN EUROPEAN COLLABORATION

BioData.pt has challenged ELIXIR Norway and ELIXIR Italy to an ELIXIR Staff Exchange Scheme that envisioned empowering National Nodes to assess the performance and impact of their activities, by adapting and adopting existing European best practices, namely from OECD, ESFRI and the RI-PATHS project.

DEMONSTRATING IMPACT

In 2020, a simple low-cost process supported by a CANVAS and dedicated guidelines was developed and applied to three BioData.pt activities: the “Ready for BioData Management?” programme, the “BioMentors Club” and the “CorkOakDB”. Preliminary results showed promising and allowed the prototype of a capacity building workshop “Assessing the Impact of Research Activities” that was successfully tested.

Funded by the ELIXIR STAFF EXCHANGE PROGRAMME and ELIXIR-CONVERGE WP4

THE "READY FOR BIODATA MANAGEMENT?" USE CASE

The "Ready for BioData Management?" capacity building programme is a valuable asset of BioData.pt that quickly achieved recognition by the Portuguese life sciences and health research & innovation community. Showcased here is the assessment of performance and impact of this programme in its first 18 months, using the referred low-cost approach. During this period, the "Ready For BioData Management?" programme delivered **15** courses to **297** trainees from **39** national and international institutions. The overall feedback indicated that the programme was of great importance for knowledge acquisition by participants. Mid- and long-term assessment will be conducted to consider the application of "Ready For BioData Management?" contents in research activities, namely, the preparation of data management plans for grant applications and others.

CERTIFIED BY

uma ação

PORTUGAL INCoDe.2030

INDUSTRY ENGAGEMENT

TRANSFERRING KNOWLEDGE TO THE INDUSTRY

BioData.pt is keen to promote the use of state-of-the-art bioinformatics tools and data management standards and best practices by the Portuguese private sector. Also, as the Portuguese node of ELIXIR, we are active members of the ELIXIR Industry focus group, where the development of new services such as infrastructure, data and bioinformatics as a service is studied and used to engage particularly with the agrofood and health industry. In 2020, a Bioinformatics in Business workshop was held in the Bioinformatics Open Days, organized by the Bioinformatics students at Universidade do Minho, to inspire the future generation of bioinformaticians applying their knowledge in the creation of new businesses or taking already existing companies to a next technological stage. Additionally, BioData.pt is collaborating with the BioMentors Club, a recently launched mentoring programme for entrepreneurs in life sciences and biotechnology, promoted by the **Portuguese Association of Bioindustry - P-BIO**.

BIOMENTORS CLUB

P-BIO, a BioData.pt partner, has recently released its **BioMentors Club**, an initiative aiming to support Portuguese projects or startups related to life sciences and biotechnology. Its mentoring programme - "Life Science and Biotechnology Mentoring Programme" - will allow the sharing of experience and knowledge through a group of experienced mentors in the area. These mentors are both people with experience in the creation of life science and biotechnology-related businesses or people with technical or scientific expertise in these areas that volunteered to share their knowledge with entrepreneurs. The BioMentors Club is now composed by 14 mentors and received 4 mentoring requests since September 2020.

PROJECTS

COLLABORATIVE FUNDING

BioData.pt - funded by PT2020

Goal: Establish the Portuguese Infrastructure for Biological Data to operate the National Node of ELIXIR. This RI is aimed to provide state of the art expertise in data management and bioinformatics, as well as computing resources to Portuguese academy and industry researchers.

Duration: 48 months Funding: 2.728.291,98€

Start Date: 17/06/2017

ELIXIR-CONVERGE - funded by H2020

Goal: Connect ELIXIR Nodes to provide FAIR data management as a service, through support, training, and the development of a data management toolkit.

Duration: 36 months Funding: 250.297,23€

Start Date: 01/02/2020

ELIXIR FUNDING

ELIXIR Implementation Studies

Overall funding: 83.091,25€

Project Plan for Interoperability Platform – Interoperability with a Purpose
Duration: 31 months
Start Date: 01/06/2019
Participation : 2 PMs

Federated Human Data
Duration: 31 months
Start Date: 01/06/2019
Participation : 1 PMs

Expanding the Galaxy: meeting the needs of ELIXIR Communities
Duration: 24 months
Start Date: 01/06/2019
Participation : 2 PMs

Deploying Reproducible Containers and Workflows Across Cloud Environments
Duration: 24 months
Start Date: 01/06/2019
Participation : 1 PMs

BioSchemas
Duration: 24 months
Start Date: 01/01/2020
Participation : 2 PMs

ELIXIR Staff Exchange Projects

Overall funding: 5.000,00€

Enabling FAIR plant phenotyping data submission through the Breeding API
Duration: 14 months
Start Date: 11/11/2019

Empowering ELIXIR Nodes to measure and communicate their performance and impact
Duration: 18 months
Start Date: 01/01/2020

Empowering ELIXIR PT capabilities to deploy and operate Local EGA instances
Duration: 12 months
Start Date: 01/01/2020

Bringing the ELIXIR plant data infrastructure to the Portuguese pulp and paper industry
Duration: 12 months
Start Date: 01/01/2020

BIODATA.PT EVENTS

Ready for BioData Management?

ITQB-NOVA

BioData.pt General Assembly

JANUARY

4th BioData.pt Technical Meeting

Forum BioData.pt: Value Creation for Healthcare System's Genomic Data

GTPB: PSLS20 - Practical Statistics for the Life Sciences

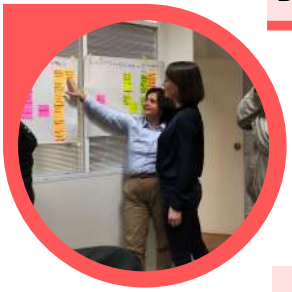
FEBRUARY

Ready for BioData Management?

IGC

ELIXIR Workshop "Impact Assessment at the National Node Level"

BioData.pt Crash Course: Microbiome Visualization with Biome-Shiny



BioData.pt Crash Course: Single-cell RNA-Seq

UMinho - BOD

Ready for BioData Management?

UMinho - BOD

BioData.pt P-BIO: Bioinformatics in Business

UMinho - BOD

COVID-19

PANDEMIC!

MARCH

JUNE

Ready for BioData Management?

iMM

Ready for BioData Management?

FCUL

@nline

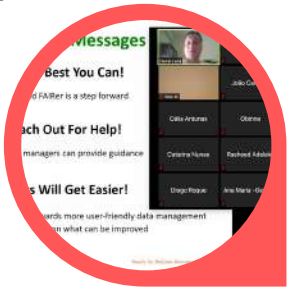
JULY

Ready for BioData Management?

UBI

Ready for BioData Management?

CCMAR



Ready for BioData Management?

UTAD

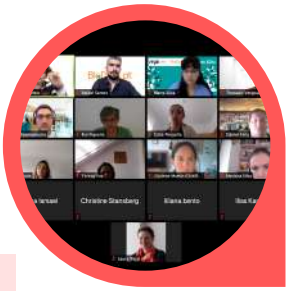
Ready for BioData Management?

IPB

Ready for BioData Management?

CNC-UC

OCTOBER

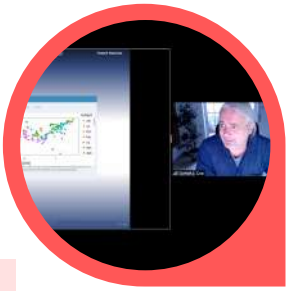


BioData.pt Workshop: Assessing the Impact of Research Activities

ELIXIR Train-the-Trainer

BioData.pt Talks: The Plant Sciences Community

NOVEMBER



Ready for BioData Management?

DMP Double Course

BioData.pt Talks: The Microbiology and Biotechnology Community

BioData.pt Talks: The Marine Resources Community

DECEMBER

COMMUNICATION

WEBSITES

Effective internal and external communications are critical to keeping internal cohesion, and reaching and engaging with BioData.pt users and stakeholders.

The **BioData.pt website** is online since 2017, to disseminate news, events, training activities, services and resources. This communication tool is under continuous improvement.

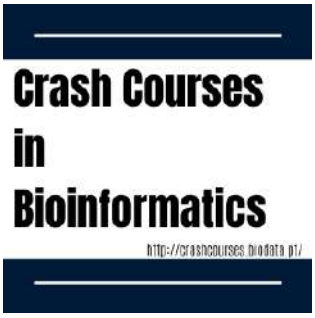
Main website

8234	20899	5588
Sessions	Page views	Users

Comparing to 2019...

+ 35,25%	+ 23,85%	+ 30,68%

BioData.pt also hosts several satellite websites with the purpose of sharing specific facets of its portfolio, namely, **BioData.pt Service Hub**, **Ready for BioData Management?**, **Impact Assessment**, and **Crash Courses in Bioinformatics**.

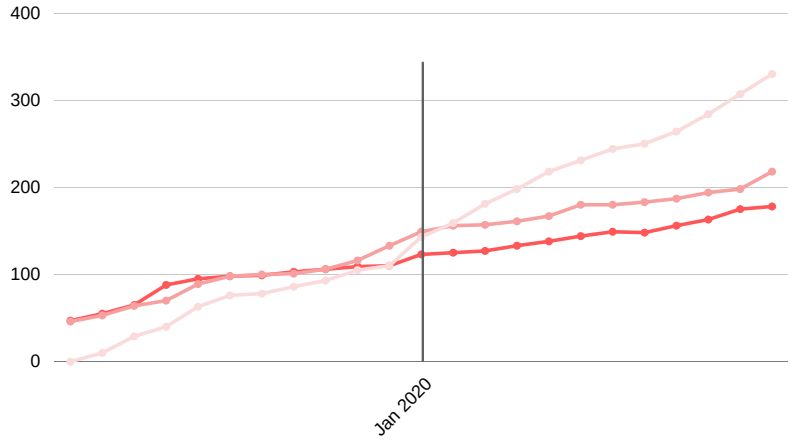


2174	4335	380	339
Page views	Page views	Page views	Page views
277	877	42	78
Users	Users	Users	Users

SOCIAL NETWORKS

BioData.pt is also present in LinkedIn, Facebook and Twitter to ensure coverage of a broader audience through community building and networking.

Followers over 2 years



> **By giving people the power to share, we're making the world more transparent.**
>
> **Mark Zuckerberg**
> **(CEO at Facebook)**

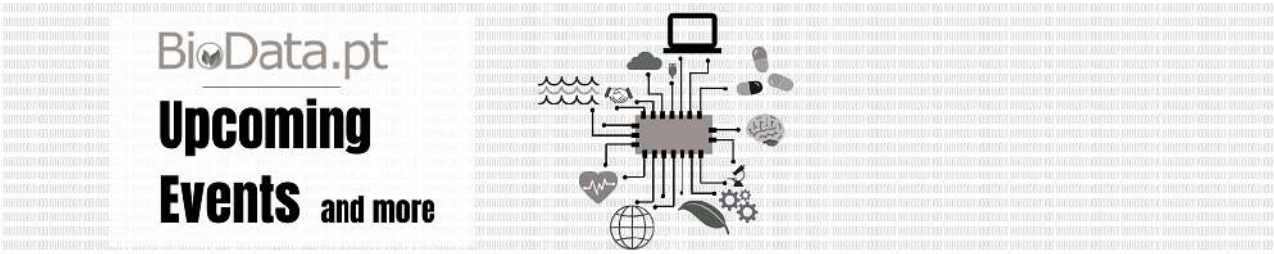
New followers in 2020

55 Facebook

85 Twitter

187 LinkedIn

THE BIODATA.PT NEWSLETTER



The **BioData.pt Newsletter** contains information about activities, events and news being both informative and enticing for our users network.

12
Newsletters

300+
People often engaged

PUBLICATIONS

- Arias-Baldrich, C., Silva, M.C., Bergeretti, F., Chaves, I., Miguel, C., Saibo, N.J., Sobral, D., Faria, D. and Barros, P.M., 2020. CorkOakDB—The Cork Oak Genome Database Portal. Database, 2020.
- Carlier, J.D., Ettamimi, S., Cox, C.J., Hammani, K., Ghazal, H. and Costa, M.C., 2020. Prokaryotic diversity in stream sediments affected by acid mine drainage. Extremophiles, 24(6), pp.809-819.
- Egger, C., Neusser, T.P., Norenburg, J., Leasi, F., Buge, B., Vannozzi, A., Cunha, R.L., Cox, C.J. and Jörger, K.M., 2020. Uncovering the shell game with barcodes: diversity of meiofaunal Caecidae snails (Truncatelloidea, Caenogastropoda) from Central America. ZooKeys, 968, p.1.
- López-Fernández, H., Vieira, C.P., Fdez-Riverola, F., Reboiro-Jato, M. and Vieira, J., 2020, June. Inferences on Mycobacterium Leprae Host Immune Response Escape and Antibiotic Resistance Using Genomic Data and GenomeFastScreen. In International Conference on Practical Applications of Computational Biology & Bioinformatics (pp. 42-50). Springer, Cham.
- Monteiro, P.T., Oliveira, J., Pais, P., Antunes, M., Palma, M., Cavalheiro, M., Galocha, M., Godinho, C.P., Martins, L.C., Bourbon, N. and Mota, M.N., 2020. YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. Nucleic acids research, 48(D1), pp.D642-D649.
- Pais, P., Galocha, M., Viana, R., Ola, M., Cavalheiro, M., Takahashi-Nakaguchi, A., Chibana, H., Butler, G. and Teixeira, M.C., 2020. Candida glabrata transcription factor Rpn4 mediates fluconazole resistance through regulation of ergosterol biosynthesis and plasma membrane permeability. Antimicrobial agents and chemotherapy, 64(9).
- Pais, P., Vagueiro, S., Mil-Homens, D., Pimenta, A.I., Viana, R., Okamoto, M., Chibana, H., Fialho, A.M. and Teixeira, M.C., 2020. A new regulator in the crossroads of oxidative stress resistance and virulence in Candida glabrata: The transcription factor CgTog1. Virulence, 11(1), pp.1522-1538.
- Papoutsoglou, E.A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I.N., Chaves, I., Coppens, F., Cornut, G., Costa, B.V., Ćwiek-Kupczyńska, H. and Droesbeke, B., 2020. Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytologist, 227(1), pp.260-273.
- Sousa, F., Civáň, P., Brazão, J., Foster, P.G. and Cox, C.J., 2020. The mitochondrial phylogeny of land plants shows support for Setaphyta under composition-heterogeneous substitution models. PeerJ, 8, p.e8995.
- Sousa, F., Civáň, P., Foster, P.G. and Cox, C.J., 2020. The chloroplast land plant phylogeny: analyses employing better-fitting tree-and site-heterogeneous composition models. Frontiers in plant science, 11, p.1062.
- Viana, R., Dias, O., Lagoa, D., Galocha, M., Rocha, I. and Teixeira, M.C., 2020. Genome-scale metabolic model of the human pathogen Candida albicans: a promising platform for drug target prediction. *Journal of Fungi*, 6(3), p.171.
- Williams, T.A., Cox, C.J., Foster, P.G., Szöllősi, G.J. and Embley, T.M., 2020. Phylogenomics provides robust support for a two-domains tree of life. Nature ecology & evolution, 4(1), pp.138-147.

PARTNERS & PEOPLE

PARTNERS

CCMAR
Centro de Ciências do Mar

inesc id
lisboa

cebal

INSTITUTO
GULBENKIAN
DE CIÊNCIA

CENTRE OF
BIOLOGICAL
ENGINEERING

TÉCNICO
LISBOA

iBB
Institute for Bioengineering and Biosciences

iTqb
UNIVERSIDADE
NOVA
DE LISBOA

iBET
Instituto de Biologia
Experimental e Tecnológica

Champalimaud
Foundation

IBMC
INSTITUTO DE BIOLOGIA MOLECULAR E CELULAR
INSTITUTE FOR MOLECULAR AND CELL BIOLOGY

P-BIO
Portugal's Biotechnology Industry Organization
Associação Portuguesa de BioIndústria

PEOPLE

José Pereira Leal
President

Mário Gaspar da Silva
Vice-President / Head of Node

Ana Portugal Melo
Executive Director / Deputy HoN

Alexandre Francisco
Technical Coordinator

Pedro Fernandes
Training Coordinator

Rafael Santos
Node Coordinator

Isabel Rocha
ELIXIR Board Member

Adelino Canário

Ana Teresa Freitas

Arlindo Oliveira

Bruno Costa

Célia Miguel

Cláudia Godinho

Cristina Vieira

Cymon Cox


Daniel Faria


Daniel Neves


Daniel Sobral


Diogo Lima


Fátima Duarte

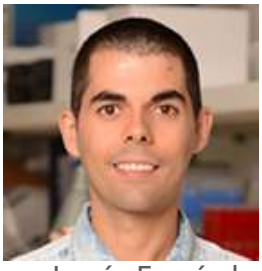

Fernando Mira da Silva

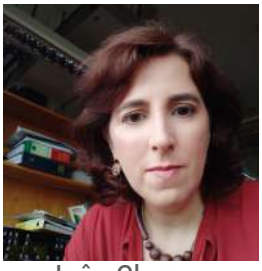

Filipa Sacadura


Filippo Bergeretti


Gian Luca di Moro


Henrique Costa


Hugo Lopéz-Fernández


Inês Chaves


Inês Costa


Inês Modesto


Isabel Sá-Correia


João Baúto


João Cardoso


João Filomena


João Garcia

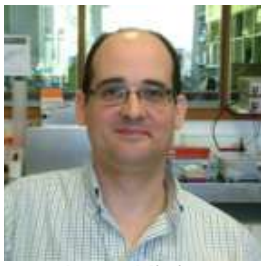

João Machado


João Raimundo


João Rato


João Sousa


Jorge Oliveira


Jorge Vieira


José Borbinha


Luís Teixeira


Mafalda Cavalheiro


Manuel Torrinha


Marcos Ramos


Maria Margarida Oliveira


Marta Silva


Miguel Cardoso


Miguel Rocha


Miguel Teixeira


Nelson Saibo


Oscar Dias


Pedro Barros


Pedro Ferreira


Pedro Garcia da Silva

Pedro Monteiro

Pedro Pais

Ricardo Leite

Rodrigo Costa

Romeu Viana

Simão Soares

Susana Val

COORDINATION

Ana Portugal Melo

CONTENTS AND DESIGN

Ana Portugal Melo, Marta Silva and Rafael Santos

PROOFREADING

Daniel Faria

BioData.pt 2020

info@biodata.pt

www.biodata.pt

Co-funded by:

FCT Fundação para a Ciência e a Tecnologia

CRESC ALGARVE 2020

Lisb@20[20]

PORTUGAL 2020

UNIÃO EUROPEIA
Fundos Europeus Estruturais e de Investimento