BioData.pt Talk

The BioData.pt Plant Sciences Community

Célia Miguel (BioData.pt / FCUL / iBET),
Pedro Barros (BioData.pt / ITQB-NOVA),
Daniel Faria (BioData.pt / INESC-ID)

October 22nd, 2020

Co-financiado por:

Overview

- Introduction
  - ELIXIR
  - ELIXIR PT | Biodata.pt
  - The Biodata Plant Sciences community
- Showcase: The Cork Oak genome portal
- Showcase: Standards & Resources for FAIR Plant Data
- Other Resources
- Conclusions

ELIXIR

Intergovernmental organisation that brings together life science resources from across Europe (databases, software tools, training materials, standards and compute resources).

Goal: coordinate life science resources from across Europe so they form a single infrastructure

This makes it easier for scientists to:

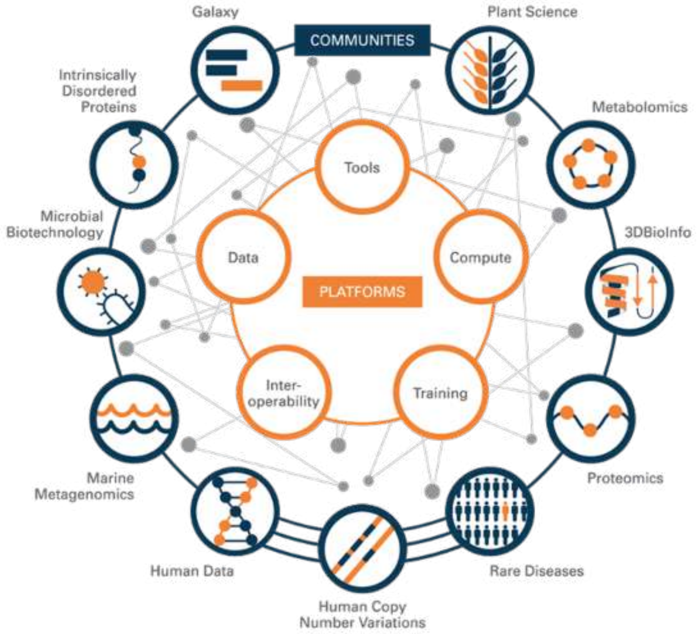
- Find and share data
- Exchange expertise
- Agree on best practices in scientific research

ELIXIR Organisation

ELIXIR coordinates activities through at least one of the five areas of activities called **Platforms**:

- Compute
- Data
- Interoperability
- Tools
- Training

These Platforms are driven by eleven ELIXIR **Communities** which develop standards, services, and training within their life science domains.

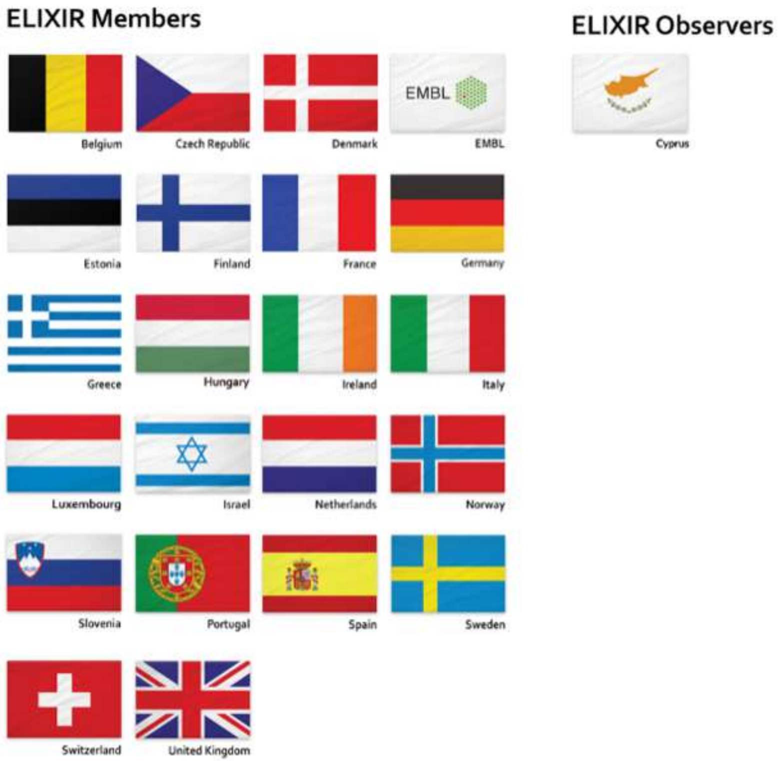


ELIXIR Members

- ELIXIR members host Nodes, which represent centres of excellence in bioinformatics.

- In total there are more than 700 researchers from over 220 research institutes.

- ELIXIR activities are coordinated by the ELIXIR Hub, based at the Wellcome Genome Campus, UK.



ELIXIR Portugal | BioData.pt

Consortium Partners

Champalimaud Foundation

inesc id lisboa

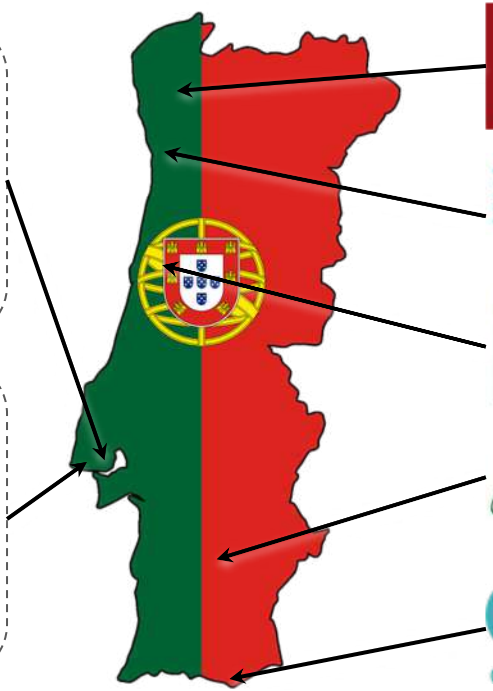
iBB Institute for Bioengineering and Biosciences

TÉCNICO LISBOA

FUNDAÇÃO CALOUSTE GULBENKIAN Instituto Gulbenkian de Ciência

iBET Instituto de Biologia Experimental e Tecnológica

ITQB INSTITUTO DE TECNOLOGIA QUÍMICA E BIOLÓGICA /UNL



UNIVERSIDADE DO MINHO

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

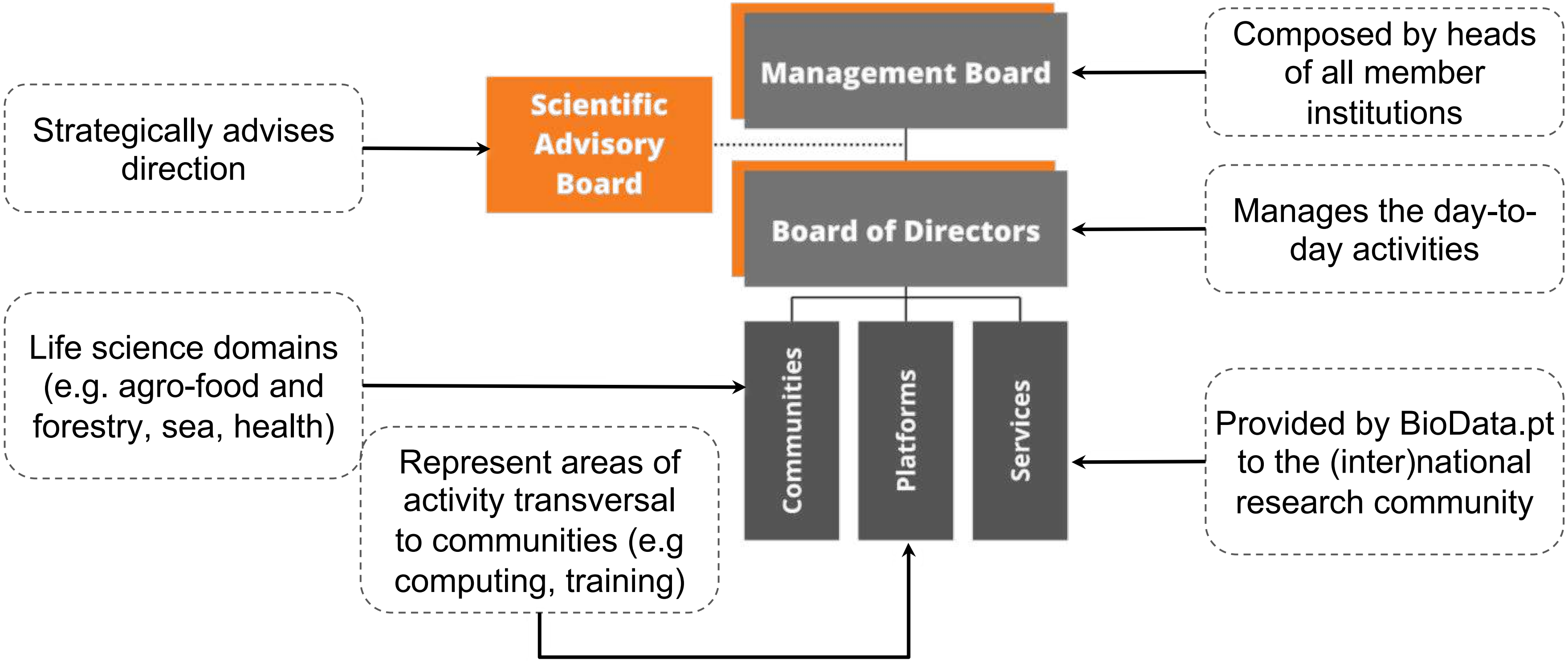
P-BIO Portugal's Biotechnology Industry Organization Associação Portuguesa de BioIndústria

cebal Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo

CCMAR Centro de Ciências do Mar

ELIXIR Portugal | BioData.pt

Organization



Strategically advises direction

Scientific Advisory Board

Management Board

Composed by heads of all member institutions

Board of Directors

Manages the day-to-day activities

Life science domains (e.g. agro-food and forestry, sea, health)

Communities

Platforms

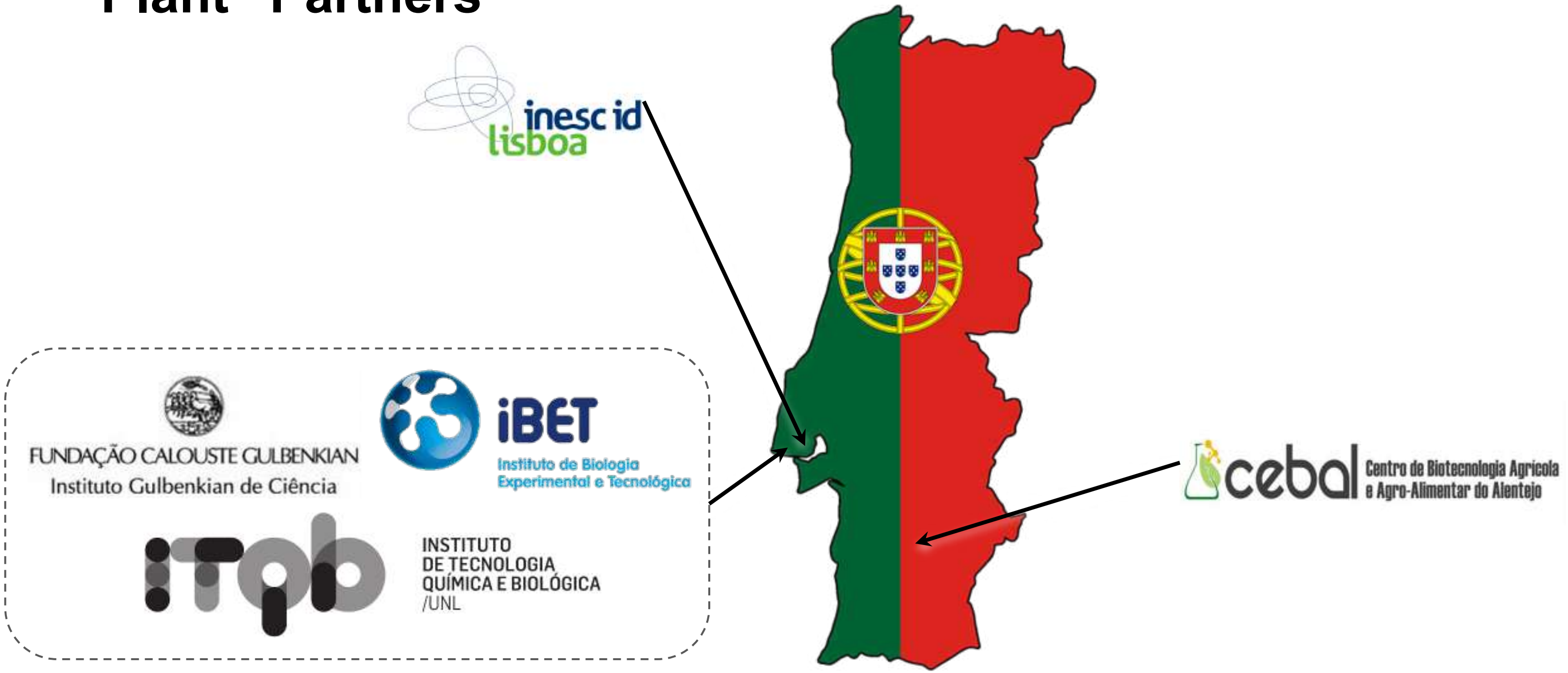
Services

Provided by BioData.pt to the (inter)national research community

Represent areas of activity transversal to communities (e.g computing, training)

The BioData.pt Plant Sciences Community

"Plant" Partners



inesc id lisboa

FUNDAÇÃO CALOUSTE GULBENKIAN
Instituto Gulbenkian de Ciência

iBET
Instituto de Biologia Experimental e Tecnológica

ITQB
INSTITUTO DE TECNOLOGIA QUÍMICA E BIOLÓGICA /UNL

cebal
Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo

The BioData.pt Plant Sciences Community

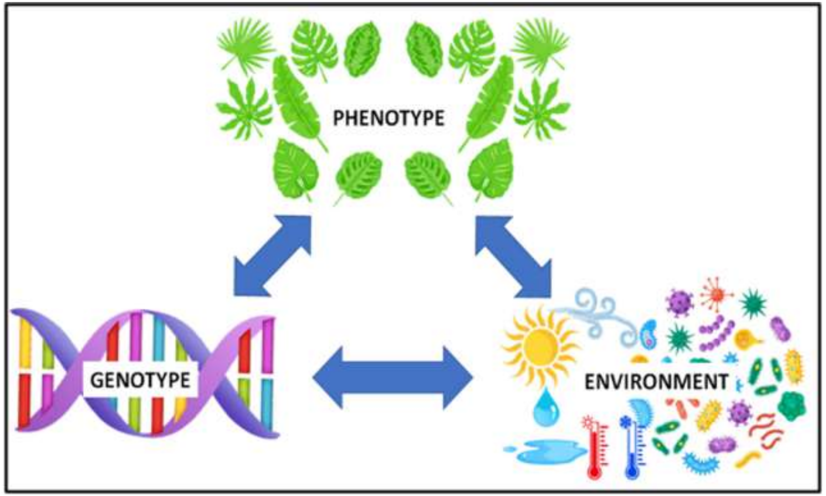
Context

- Became active during the ELIXIR-EXCELERATE project, where we co-led the Plant Sciences use-case
- **Goal**: support the publication and access to phenotypic and genotypic data, annotated according to established standards (woody plants as the main target)
  - Critical for addressing some of the major global challenges:
    - Sustainable supply of food and non-food materials
    - Competitive life-sciences industry sector
    - Environmental protection

The BioData.pt Plant Sciences Community

Motivation: Plant Breeding (GxExP) remains challenging

- Advances can be obtained from the integration of genomic/genotyping data with diverse types of phenotyping data
- Systematic study of phenotypes on a genome-wide scale, and its association with genomic information under different environmental conditions



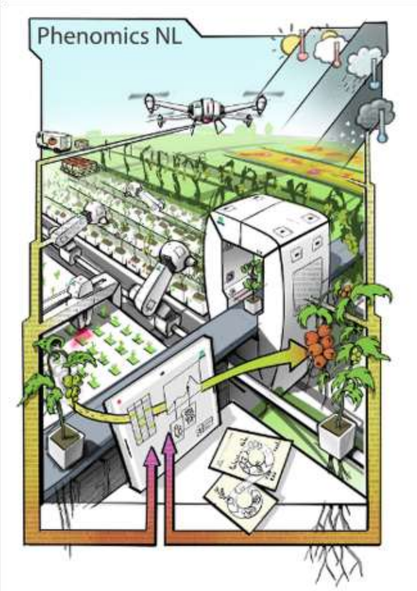
Djande et al 2020 DOI 10.3390/agronomy10060831

- Genomics/genotyping and phenotyping datasets are growing in number and size ⇒ gaps in genotype-phenotype associations

The BioData.pt Plant Sciences Community

Motivation: How to Make Plant Data FAIR?

- Problems with plant data:
  - Heterogeneous (different settings, types of data...)
  - Complex and diverse experimental designs
  - Dispersed (no comprehensive public archive)
  - Poorly annotated (weaknesses in standards)
- FAIR data principles (Findable, Accessible, Interoperable and Reusable) are key to enable knowledge discovery



The BioData.pt Plant Sciences Community

Objectives & Activities

Develop/recommend standards and ontologies to enable FAIR plant phenotyping data

Develop/implement repositories for plant phenotyping and genomic data

Develop/implement user-friendly interfaces for data deposition and retrieval

Icons made by inipagistudio, phatplus, Eucalyp and Freepik

Annotate and provide curated plant data sets

Develop/provide tools for plant data analysis

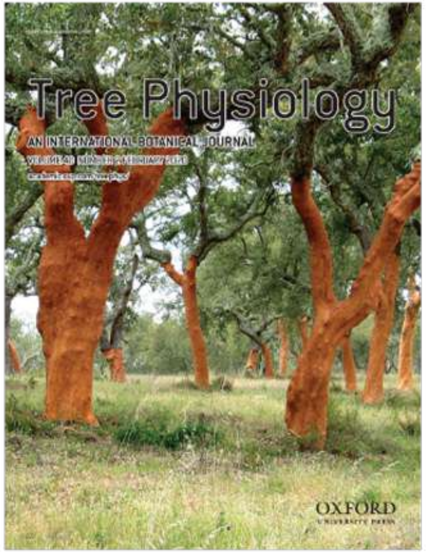
Provide training on plant data management

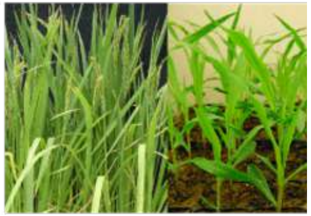
Engage with industry to exchange and apply knowledge

The BioData.pt Plant Sciences Community

Species of Interest

- Cork oak (*Q. suber*)
- Maritime Pine (*P. pinaster*)
- Eucalyptus (*E. globulus*)
- Olive tree (*O. europaea*)
- Grapevine (*V. vinifera*)
- Rice (*O. sativa*)
- ...



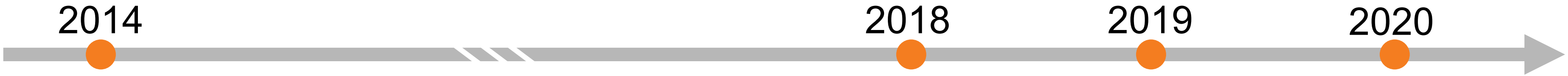


Plant Sciences Community Showcase

CorkOakDB: The Cork Oak Genome Portal

CorkOakDB: The Cork Oak Genome Portal

History

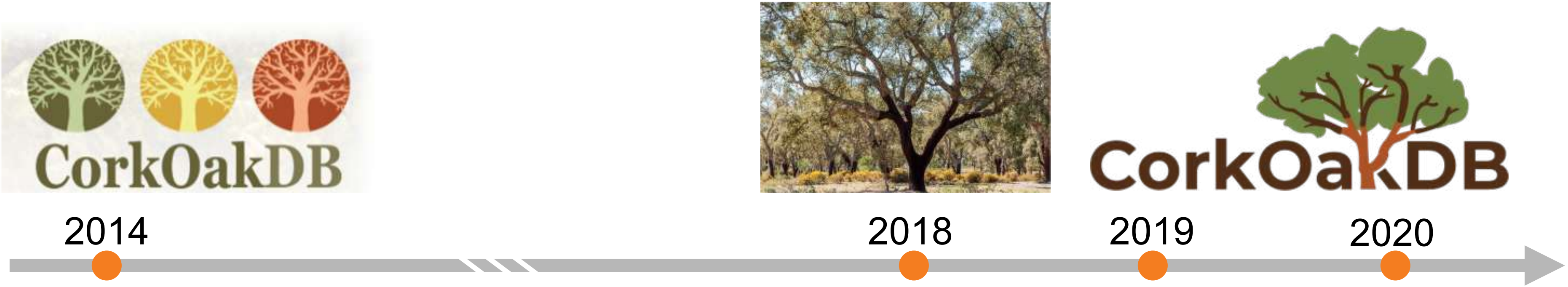


- Comprehensive database of cork oak transcriptome obtained by cDNA sequencing (ESTs)
- No knowledge of genome structure
- Incomplete gene coding sequences

www.corkoakdb.org

CorkOakDB: The Cork Oak Genome Portal

History



2014

2018

2019

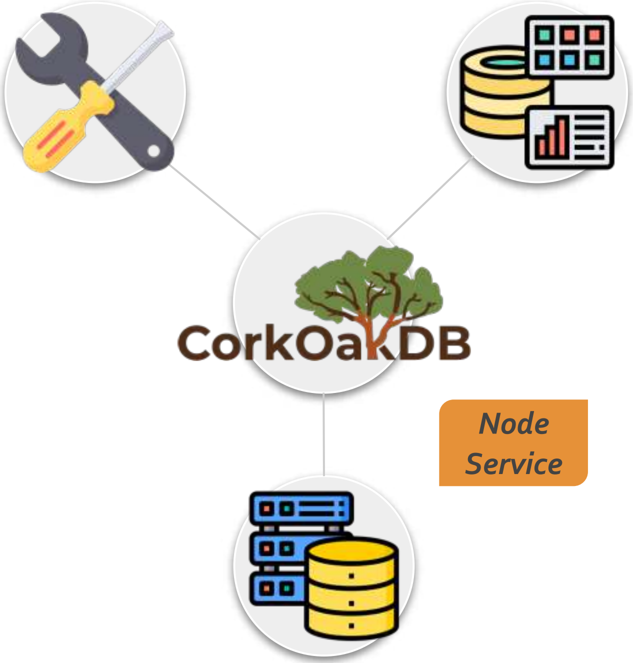
2020

- Reference genome sequenced
- Improved gene structure annotation
- New transcriptomic datasets publicly available

CorkOakDB: The Cork Oak Genome Portal

Goal

- Create an integrated repository dedicated to cork oak ‘omics’
- Development of tools for data visualization and retrieval for core genomics analyses
- Become a reference hub for research, aggregating all available genomic and transcriptomic (...) data



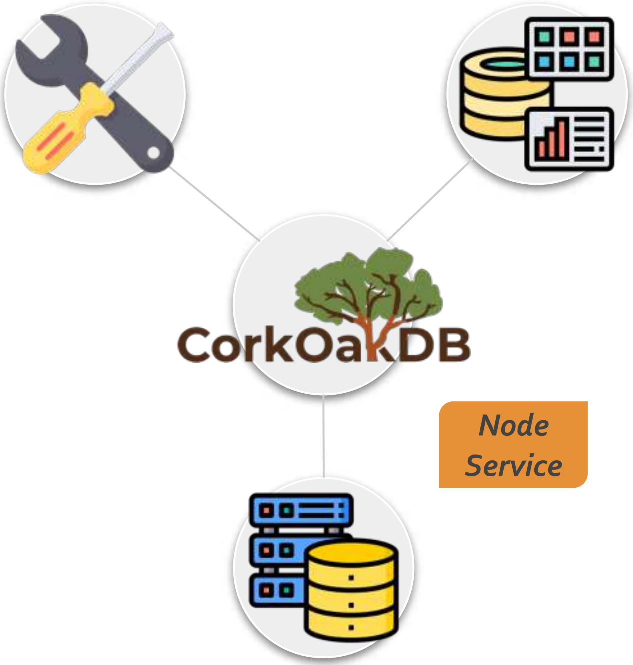
CorkOakDB: The Cork Oak Genome Portal

Data

- Genome sequence and structural annotation (gene, exon, intron, CDS)
- Gene expression (retrieved from publicly available studies)
- Curated metadata

Structure

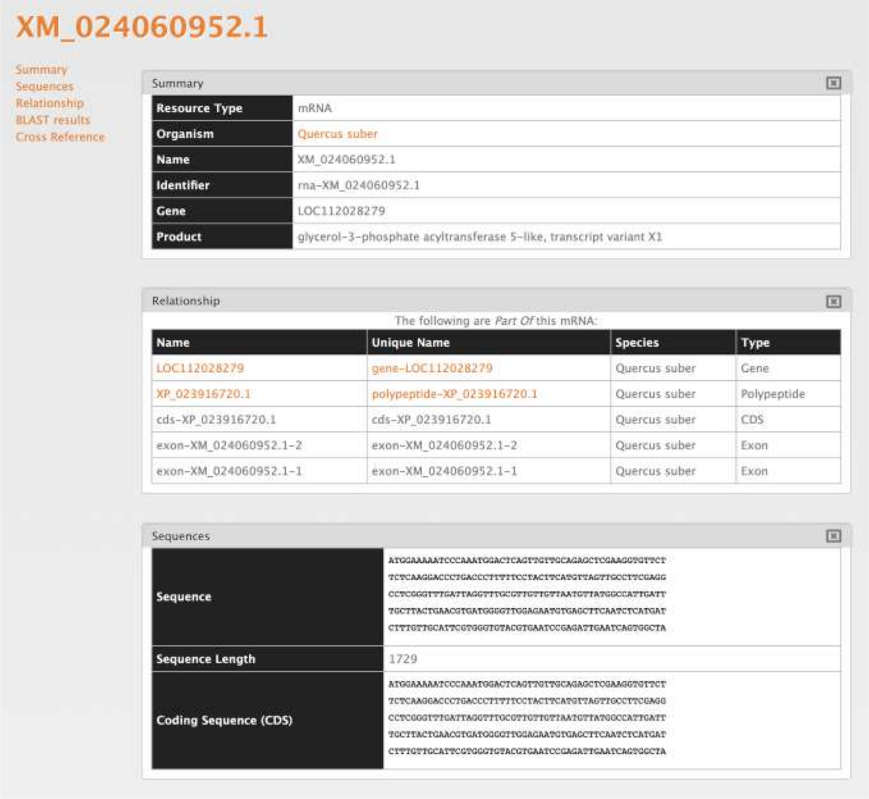
- Web portal based on the Tripal framework (Drupal CMS + Chado database)



CorkOakDB: Genome Browser

Tools

- **Gene search and sequence retrieval**
- Genome visualization (JBrowse)
- Homology search (Blast)



XM_024060952.1

Summary
Sequences
Relationship
BLAST results
Cross Reference

Summary

Resource Type	mRNA
Organism	Quercus suber
Name	XM_024060952.1
Identifier	rna-XM_024060952.1
Gene	LOC112028279
Product	glycerol-3-phosphate acyltransferase 5-like, transcript variant X1

Relationship

The following are *Part Of* this mRNA:

Name	Unique Name	Species	Type
LOC112028279	gene-LOC112028279	Quercus suber	Gene
XP_023916720.1	polypeptide-XP_023916720.1	Quercus suber	Polypeptide
cds-XP_023916720.1	cds-XP_023916720.1	Quercus suber	CDS
exon-XM_024060952.1-2	exon-XM_024060952.1-2	Quercus suber	Exon
exon-XM_024060952.1-1	exon-XM_024060952.1-1	Quercus suber	Exon

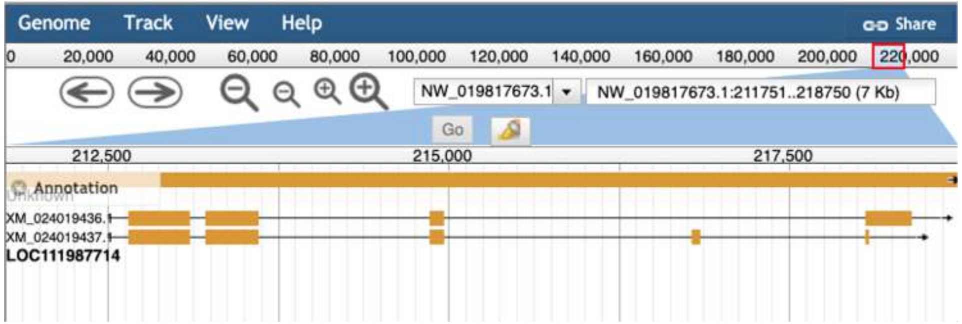
Sequences

Sequence	ATGGAAAAATCCCAAATGGACTCAGTTGTTGCAGAGCTCGAAGGTGTTCT TCTCAAGGACCCTGACCCTTTTTCCTACTTCATGTTAGTTGCCTTCGAGG CCTCGGGTTTGATTAGGTTTGCGTTGTTGTTAATGTTATGGCCATTGATT TGCTTACTGAACGTGATGGGGTTGGAGAATGTGAGCTTCAATCTCATGAT CTTTGTTGCATTCGTGGGTGTACGTGAATCCGAGATTGAATCAGTGGCTA
Sequence Length	1729
Coding Sequence (CDS)	ATGGAAAAATCCCAAATGGACTCAGTTGTTGCAGAGCTCGAAGGTGTTCT TCTCAAGGACCCTGACCCTTTTTCCTACTTCATGTTAGTTGCCTTCGAGG CCTCGGGTTTGATTAGGTTTGCGTTGTTGTTAATGTTATGGCCATTGATT TGCTTACTGAACGTGATGGGGTTGGAGAATGTGAGCTTCAATCTCATGAT CTTTGTTGCATTCGTGGGTGTACGTGAATCCGAGATTGAATCAGTGGCTA

CorkOakDB: Genome Browser

Tools

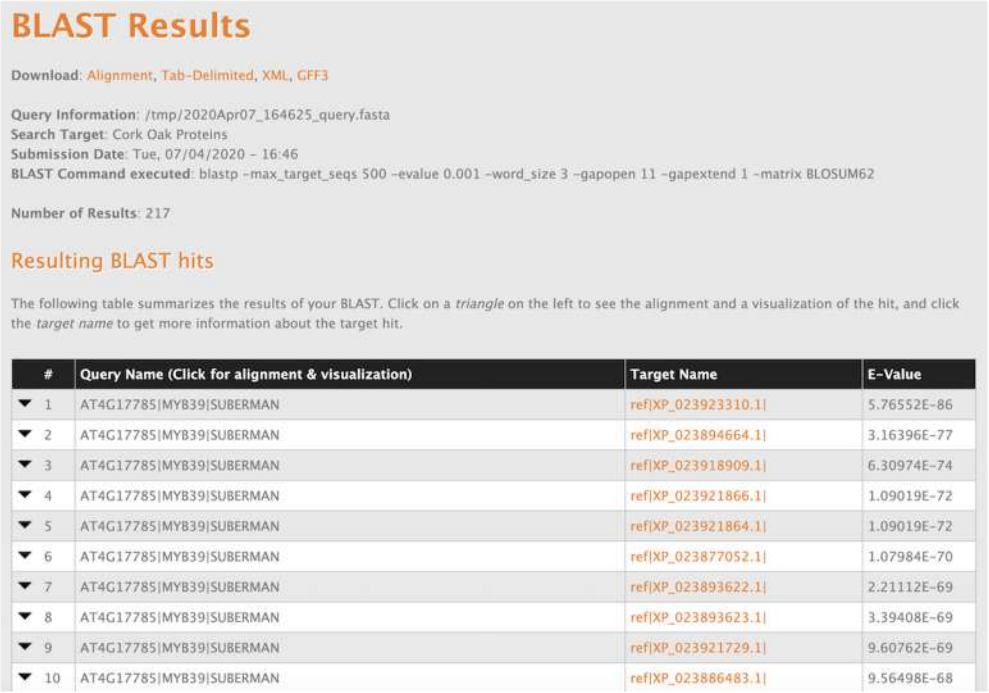
- Gene search and sequence retrieval
- **Genome visualization (JBrowse)**
- Homology search (Blast)



CorkOakDB: Genome Browser

Tools

- Gene search and sequence retrieval
- Genome visualization (JBrowse)
- **Homology search (Blast)**



BLAST Results

Download: Alignment, Tab-Delimited, XML, GFF3

Query Information: /tmp/2020Apr07_164625_query.fasta
Search Target: Cork Oak Proteins
Submission Date: Tue, 07/04/2020 - 16:46
BLAST Command executed: blastp -max_target_seqs 500 -evalue 0.001 -word_size 3 -gapopen 11 -gapextend 1 -matrix BLOSUM62

Number of Results: 217

Resulting BLAST hits

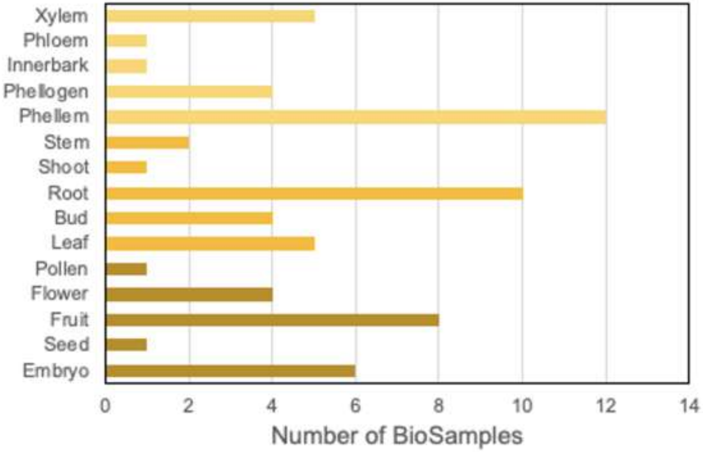
The following table summarizes the results of your BLAST. Click on a *triangle* on the left to see the alignment and a visualization of the hit, and click the *target name* to get more information about the target hit.

#	Query Name (Click for alignment & visualization)	Target Name	E-Value
▼ 1	AT4G17785|MYB39|SUBERMAN	ref|XP_023923310.1|	5.76552E-86
▼ 2	AT4G17785|MYB39|SUBERMAN	ref|XP_023894664.1|	3.16396E-77
▼ 3	AT4G17785|MYB39|SUBERMAN	ref|XP_023918909.1|	6.30974E-74
▼ 4	AT4G17785|MYB39|SUBERMAN	ref|XP_023921866.1|	1.09019E-72
▼ 5	AT4G17785|MYB39|SUBERMAN	ref|XP_023921864.1|	1.09019E-72
▼ 6	AT4G17785|MYB39|SUBERMAN	ref|XP_023877052.1|	1.07984E-70
▼ 7	AT4G17785|MYB39|SUBERMAN	ref|XP_023893622.1|	2.21112E-69
▼ 8	AT4G17785|MYB39|SUBERMAN	ref|XP_023893623.1|	3.39408E-69
▼ 9	AT4G17785|MYB39|SUBERMAN	ref|XP_023921729.1|	9.60762E-69
▼ 10	AT4G17785|MYB39|SUBERMAN	ref|XP_023886483.1|	9.56498E-68

CorkOakDB: Gene Expression

Data

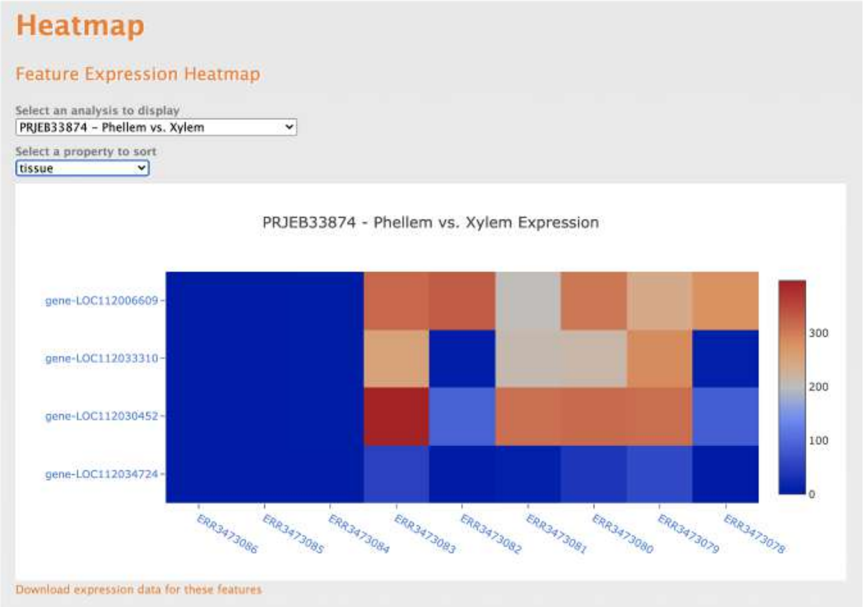
- EST sequencing data
- Transcriptomic data for different **tissues**, **developmental stages** and **growth conditions**
- 65 RNA-seq libraries from 15 studies publicly available



CorkOakDB: Gene Expression

Visualization

- RNA-seq reads aligned with the reference genome to **estimate gene expression**
- Heatmap for **comparative analysis** of multiple genes



Heatmap

Feature Expression Heatmap

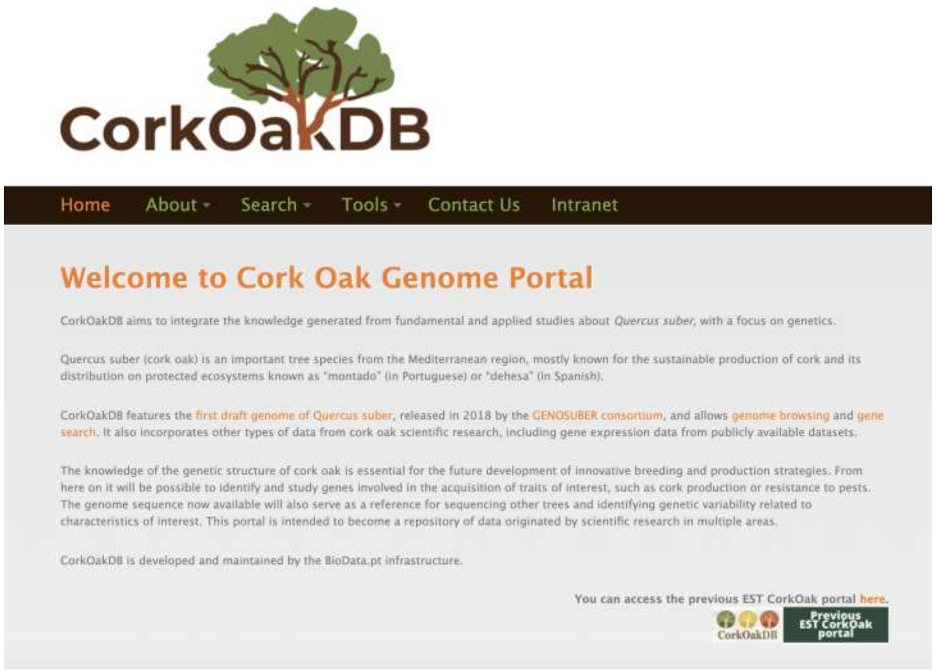
Select an analysis to display: PRJEB33874 - Phellem vs. Xylem

Select a property to sort: tissue

PRJEB33874 - Phellem vs. Xylem Expression

Download expression data for these features

CorkOakDB: Next Steps



Welcome to Cork Oak Genome Portal

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

Quercus suber (cork oak) is an important tree species from the Mediterranean region, mostly known for the sustainable production of cork and its distribution on protected ecosystems known as "montado" (in Portuguese) or "dehesa" (in Spanish).

CorkOakDB features the first draft genome of Quercus suber, released in 2018 by the GENOSUBER consortium, and allows genome browsing and gene search. It also incorporates other types of data from cork oak scientific research, including gene expression data from publicly available datasets.

The knowledge of the genetic structure of cork oak is essential for the future development of innovative breeding and production strategies. From here on it will be possible to identify and study genes involved in the acquisition of traits of interest, such as cork production or resistance to pests. The genome sequence now available will also serve as a reference for sequencing other trees and identifying genetic variability related to characteristics of interest. This portal is intended to become a repository of data originated by scientific research in multiple areas.

CorkOakDB is developed and maintained by the BioData.pt infrastructure.

You can access the previous EST CorkOak portal here.

www.corkoakdb.org

- Linking ESTs IDs (old portal) with gene IDs (new portal)
- Integration of curated data from published results (functional validation)
- Integration of other types of data (SNPs, phenotype, …)

Plant Sciences Community Showcase

Standards & Resources for FAIR Plant Data

The FAIR Data Principles

Findability

- Persistent identifiers
- Rich metadata
- Searchable repository

Interoperability

- Knowledge representation language
- Controlled vocabularies

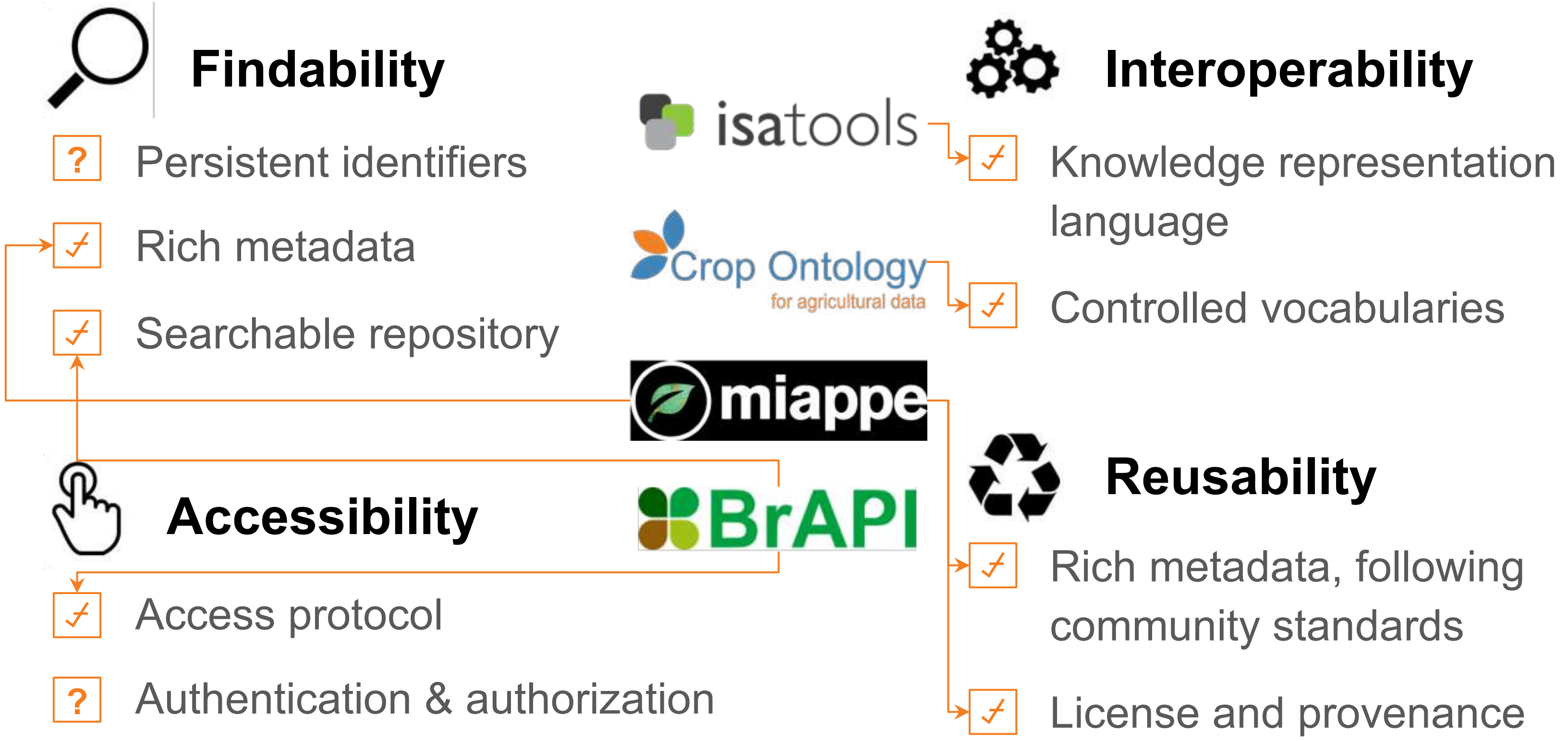
Accessibility

- Access protocol
- Authentication & authorization

Reusability

- Rich metadata, following community standards
- License and provenance

Plant Phenotyping Data FAIRness?



Findability

- ? Persistent identifiers
- ✓ Rich metadata
- ✓ Searchable repository

Accessibility

- ✓ Access protocol
- ? Authentication & authorization

Interoperability

- ✓ Knowledge representation language
- ✓ Controlled vocabularies

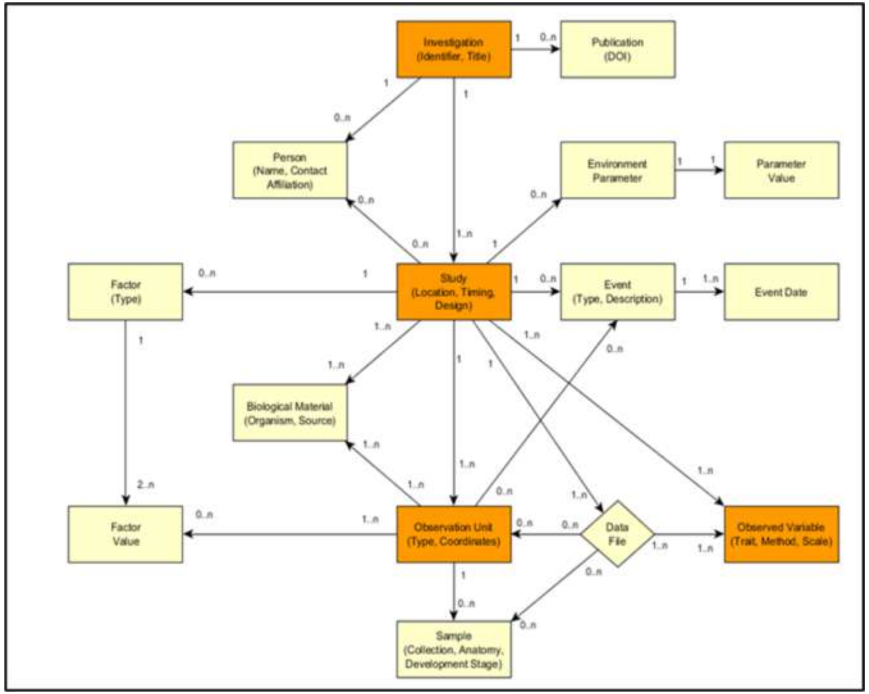
Reusability

- ✓ Rich metadata, following community standards
- ✓ License and provenance

Updating MIAPPE

Minimum Information About a Plant Phenotyping Experiment 1.1

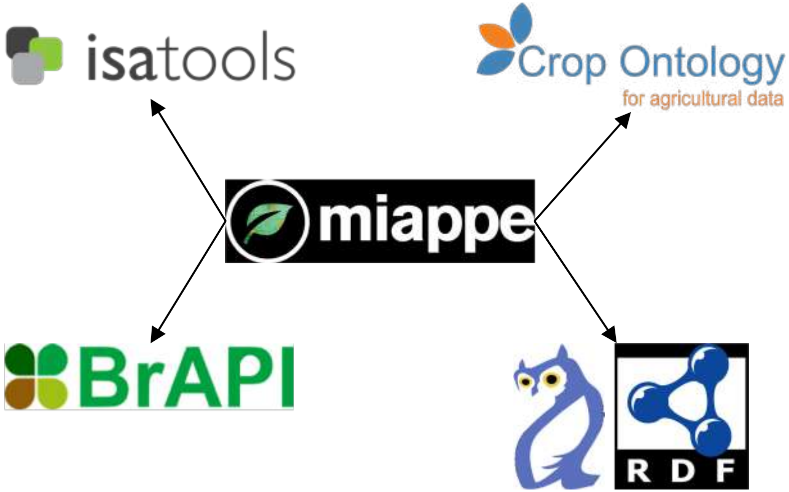
- Scope extension: woody plants (e.g. identification through GPS)
- Revised structure to match ISA
- Enriched examples
- Data model specification (a standard is more than a flat list)



Updating MIAPPE

Minimum Information About a Plant Phenotyping Experiment 1.1

- Explicit ontology recommendations
- Mapping to BrAPI
- ISA-Tab templates
- OWL encoding (PPEO) to enable RDF
- JSON-schema version (in progress)



Developing Ontologies for Plant Data Annotation

Crop Ontologies

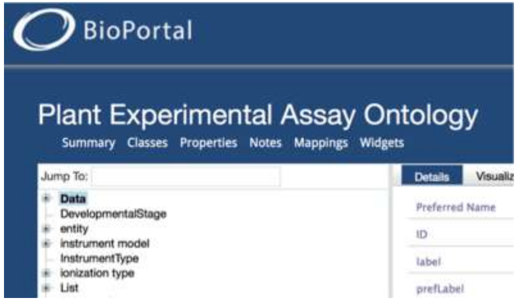
AgroPortal LIRMM

Crop Ontology for agricultural data

- Describe traits/features and methods for measuring them in specific plant species (MIAPPE observed variables)
    - ○ Woody Plant Ontology
    - ○ Rice Ontology

Plant Experimental Assay Ontology

Node Service

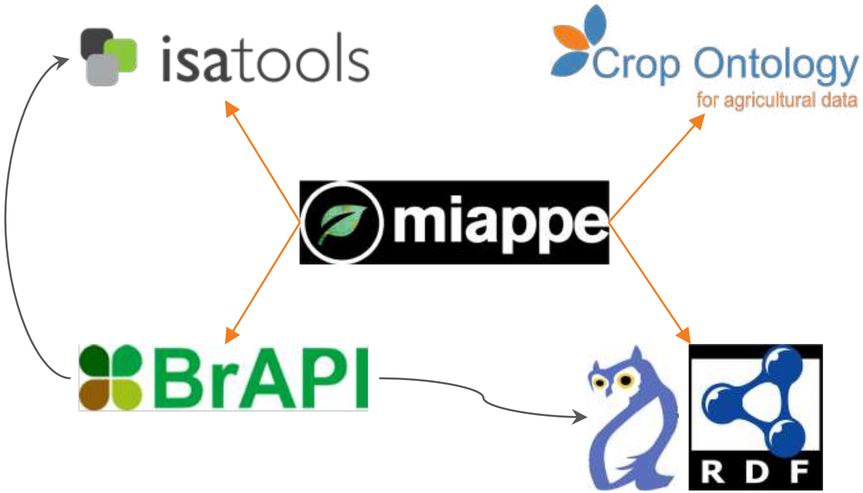


- Describes experimental procedures and pipelines in Plant (Molecular) Biology

Updating BrAPI

Breeding API 2.0

- Full coverage of mandatory MIAPPE fields (based on mapping)
- BrAPI to ISA-Tab exporting
- BrAPI to RDF or JSON-LD exporting (using PPEO as context)

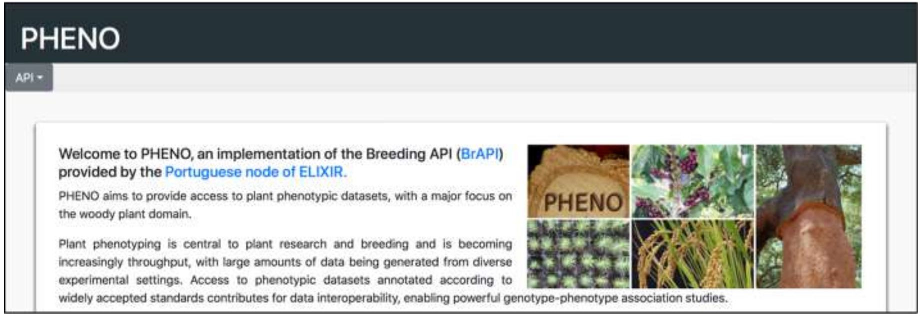


Setting up a national BrAPI end-point

Breeding API

- BrAPI web service implements 13 BrAPI calls

- Underlying SQL database reverse-engineered from BrAPI specifications (shifting to RDF)



Welcome to PHENO, an implementation of the Breeding API (BrAPI) provided by the Portuguese node of ELIXIR.

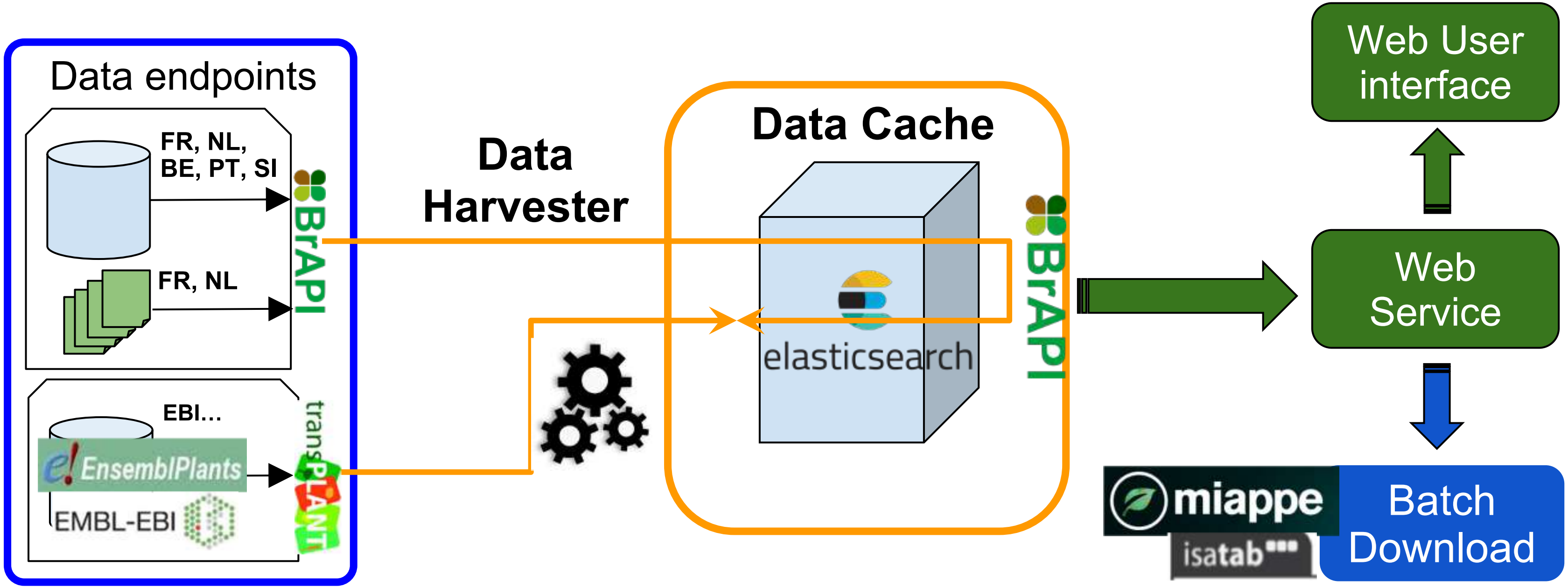
PHENO aims to provide access to plant phenotypic datasets, with a major focus on the woody plant domain.

Plant phenotyping is central to plant research and breeding and is becoming increasingly throughput, with large amounts of data being generated from diverse experimental settings. Access to phenotypic datasets annotated according to widely accepted standards contributes for data interoperability, enabling powerful genotype-phenotype association studies.

- Currently includes manually curated datasets on cork oak (Q. suber), rice (O. sativa) and J. curcas

Setting up a federated data lookup service

FAIR Data-finder for Agronomic REsearch



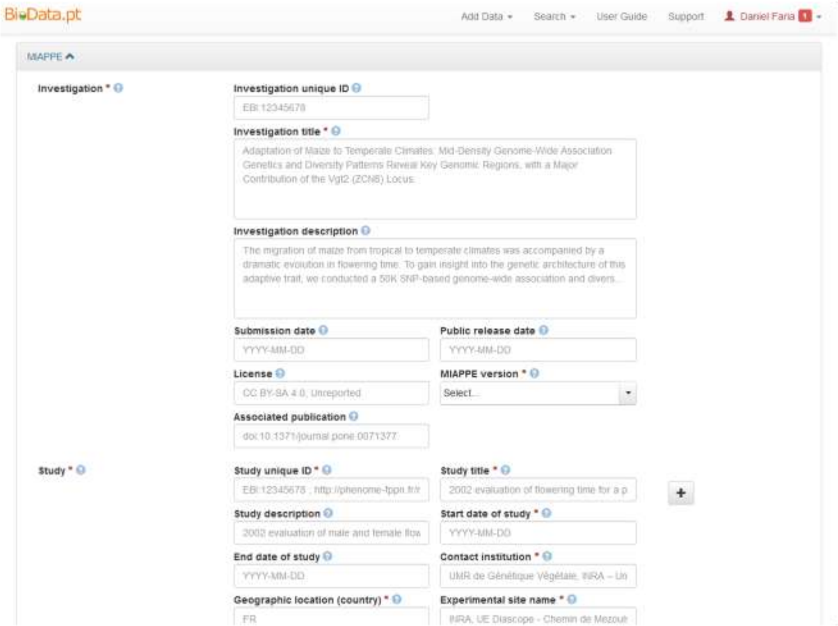
Enabling Web Page Findability with BioSchemas

- BioSchemas:
  - Extension of schema.org for the life science domain
  - Mark-up of web pages to enable findability
- Goals:
  - Map BioSchemas to MIAPPE fields critical for dataset findability
  - Mark-up CorkOakDB pages with BioSchemas
  - Setup FAIDARE to ingest BioSchemas mark-up

Improving MIAPPE Usability

User Interfaces for MIAPPE Dataset Submission

- ISA-Tab template for ISA-Tools, not the most user-friendly
- Templates for popular data management platforms:
  - Dataverse (ready for use)
  - SEEK 4 Science (nearly finalized)
- Stand-alone OWL-based interface (in progress)



Investigation *

Investigation unique ID: EBI:12345678

Investigation title *: Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with a Major Contribution of the Vgt2 (ZCN8) Locus.

Investigation description: The migration of maize from tropical to temperate climates was accompanied by a dramatic evolution in flowering time. To gain insight into the genetic architecture of this adaptive trait, we conducted a 50K SNP-based genome-wide association and divers...

Submission date: YYYY-MM-DD

Public release date: YYYY-MM-DD

License: CC BY-SA 4.0, Unreported

MIAPPE version *: Select...

Associated publication: doi:10.1371/journal.pone.0071377

Study *

Study unique ID *: EBI:12345678 ; http://phenome-fppn.fr/r

Study title *: 2002 evaluation of flowering time for a p

Study description: 2002 evaluation of male and female flow

Start date of study *: YYYY-MM-DD

End date of study: YYYY-MM-DD

Contact institution *: UMR de Génétique Végétale, INRA – Un

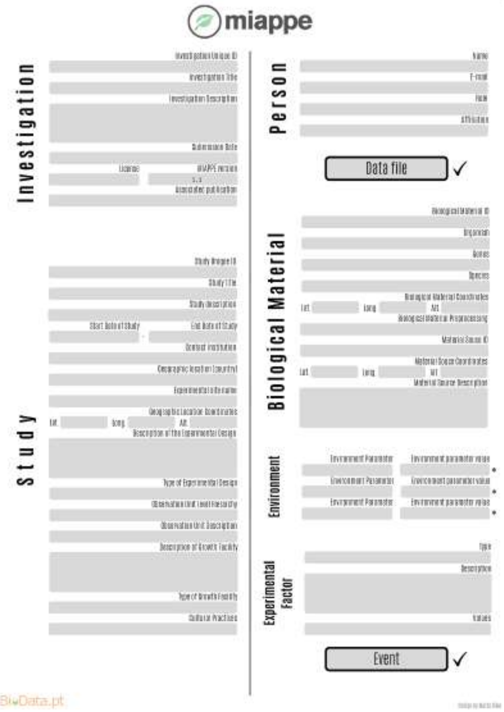
Geographic location (country) *: FR

Experimental site name *: INRA, UE Diascope - Chemin de Mezouil

Providing Training on MIAPPE

Plant Data Management Workshops

- Goal: Teach participants to describe a dataset according to MIAPPE v1.1
- Methodology: Lectures on introductory data management and MIAPPE, followed by a hands-on group exercise on a mock dataset, using a prepared MIAPPE template
- Past Events: Oeiras, Nov-2019; Paris, Feb-2020
- Future Events: TBD, virtual!



miappe

Investigation

Investigation Unique ID

Investigation Title

Investigation Description

Submission Date

License

MIAPPE version

1.1

Associated publication

Study

Study Unique ID

Study Title

Study description

Start Date of Study

End Date of Study

Contact institution

Geographic location (country)

Experimental site name

Geographic Location Coordinates

Lat Long Alt

Description of the Experimental Design

Type of Experimental Design

Observation Unit Level Hierarchy

Observation Unit Description

Description of Growth Facility

Type of Growth Facility

Cultural Practices

Person

Name

E-mail

Role

Affiliation

Data file ✓

Biological Material

Biological Material ID

Organism

Genus

Species

Biological Material Coordinates

Lat Long Alt

Biological Material Preprocessing

Material Source ID

Material Source Coordinates

Lat Long Alt

Material Source Description

Environment

Environment Parameter | Environment parameter value

Environment Parameter | Environment parameter value

Environment Parameter | Environment parameter value

Experimental Factor

Type

Description

Values

Event ✓

BioData.pt

Engaging with the Industry

Data Producers: The Navigator Company

- A major player in the international pulp and paper market and one of Portugal's strongest brands on the world stage
- Massive amounts of data on eucalyptus breeding and genetics:
    - Genotypic and phenotypic data on over 300,000 specimens across a range of sites and covering up to 4 generations of pedigree

THE NAVIGATOR COMPANY

WOOD
CELLULOSE
ENVIRONMENT
OTHER USES
SOCIO-ECONOMIC IMPACT

Engaging with the Industry

Data Producers: The Navigator Company

Goals:

- Process, annotate and ingest pilot datasets into our BrAPI end-point
- Share knowledge and demonstrate the value of plant data FAIRification (e.g. integration with external datasets)
- Automate the process so the company can submit its datasets to BrAPI in bulk

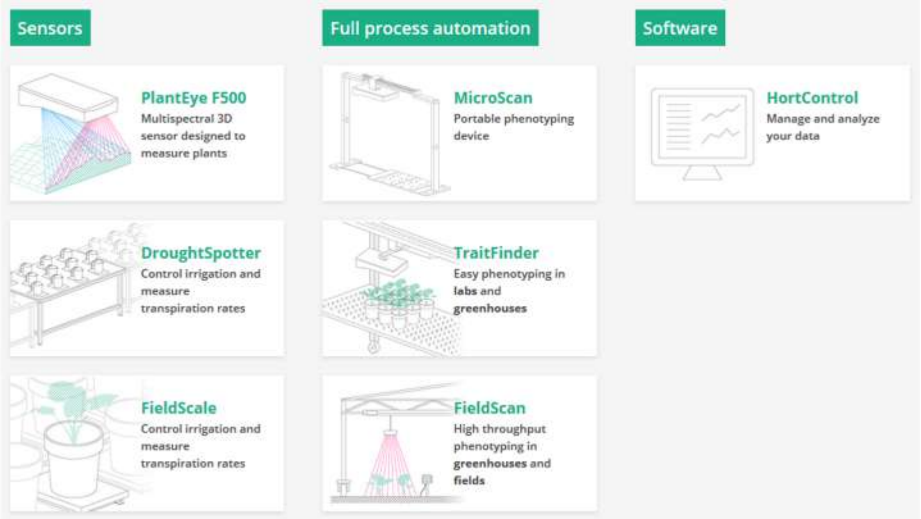
THE NAVIGATOR COMPANY

Engaging with the Industry

Technology Producers: Phenospex

- Provide state-of-the-art automated plant phenotyping solutions to researchers across the world
- Software team interested in ensuring their information system is MIAPPE compliant and can export data via BrAPI calls
- An ideal partnership to ensure researchers produce FAIR data!

PHENOSPEX
Smart Plant Analysis



Sensors

PlantEye F500
Multispectral 3D sensor designed to measure plants

DroughtSpotter
Control irrigation and measure transpiration rates

FieldScale
Control irrigation and measure transpiration rates

Full process automation

MicroScan
Portable phenotyping device

TraitFinder
Easy phenotyping in labs and greenhouses

FieldScan
High throughput phenotyping in greenhouses and fields

Software

HortControl
Manage and analyze your data

Putting it All Together

One Standard to Rule Them All

- Whichever data management tool or data submission platform you choose for plant phenotyping data shall be MIAPPE compliant
- Metadata from all tools and platforms shall be interchangeable
- Genotyping and genomic data shall comply with MIAPPE specifications of Biological Materials to enable integration
- Bioschemas shall map to upper layers of MIAPPE

miappe

isatools

The Dataverse® Project

SEEK

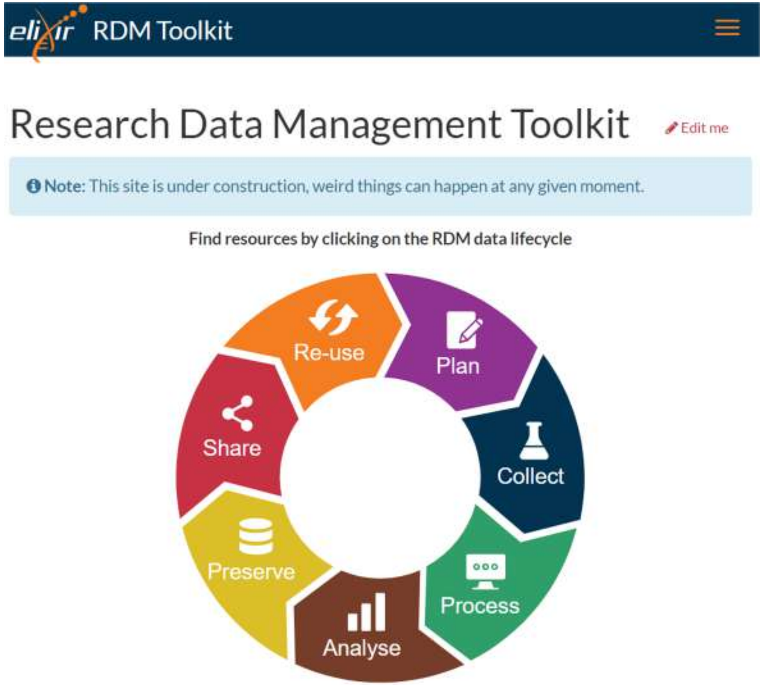
BrAPI

RDF

Putting it All Together

The ELIXIR RDM Toolkit – Plant Domain

- Document tools and resources for plant FAIR data management
- Address FAQs and common needs
- Prepare Data Management Plan templates for the plant domain, referencing the resources developed by the community and listed in the RDM Toolkit



elixir RDM Toolkit

Research Data Management Toolkit ✎Edit me

ℹ Note: This site is under construction, weird things can happen at any given moment.

Find resources by clicking on the RDM data lifecycle

Other Resources

Conclusions

Plant sRNA Portal

miRPursuit

- Automated workflow for downstream analysis of gene expression data and prediction of sRNA target genes

sRNA Database

- Repository of sRNA either annotated using miRPursuit or publicly available

Node Service

Plant sRNA Portal

Under Development!

Conclusions & Perspectives

Engagement and adoption of resources by the community to promote FAIR data and provide easier ways to analyse and gain new insights from available data…

Training sessions on the usage of resources, plant data management, training materials…

Icons made by Freepik

Collaboration with other ELIXIR nodes - European projects (Converge,…), Implementation Studies, ELIXIR Knowledge Exchange and Staff Exchange Programs,…

Interaction with EMPHASIS, Crop Ontology, Bioversity,…

Thanks!

The BioData.pt Plant Sciences Community

Célia Miguel, FCUL / iBET
Margarida Oliveira, ITQB-NOVA
Nelson Saibo, ITQB-NOVA
Pedro Barros, ITQB-NOVA
Marcos Ramos, CEBAL
Inês Chaves, ITQB-NOVA
Daniel Faria, INESC-ID / IGC
Bruno Costa, FCUL / INESC-ID
Marta Silva, ITQB-NOVA
Filippo Bergeretti, ITQB-NOVA
André Cordeiro, ITQB-NOVA

Past collaborators:
Daniel Sobral
Cirenia Arias-Baldrich

All the elixir Plant Sciences Community

Especially our close collaborators:
Cyril Pommier, ELIXIR-FR
Anne-Françoise Adam-Blondon, ELIXIR-FR
Celia Michotey, ELIXIR-FR
Richard Finkers, ELIXIR-NL
Evangelia Papoutsoglou, ELIXIR-NL
Frederik Coppens, ELIXIR-BE
Paul Kersey, previously ELIXIR-EBI

elixir PORTUGAL BioData.pt

Co-financiado por:
CRESC ALGARVE 2020 Lisb@20²⁰ PORTUGAL 2020 UNIÃO EUROPEIA Fundos Europeus Estruturais e de Investimento