BioData.pt elixir PORTUGAL

Ready for BioData Management?

DAY 2 DAY DATA MANAGEMENT COURSE

FOR LIFE SCIENCES

Funding:

- ELIXIR Implementation Study "Impact evaluation at Node-level - getting it done“
- ELIXIR-CONVERGE Project “Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services”

BioData.pt elixir PORTUGAL

Ready for BioData Management?

DAY 2 DAY DATA MANAGEMENT COURSE

FOR LIFE SCIENCES

IGC USE CASE: FROM PROBLEMS TO SOLUTIONS

BY CAROLINA VENTURA-COSTA, IGC

GULBENKIAN CIÊNCIA

QUANTITATIVE & DIGITAL SCIENCE UNIT

The Quantitative Biology & Digital Science Unit supports researchers in quantifying, analyzing and explaining their data. The Unit also provides services in data management and reproducible workflows.

Biology has entered the big data age. Now, more than ever, computational and analytical skills are required to properly interrogate and interpret data. We provide expertise on data analysis, mathematical modelling, and statistical consulting and offer services on the whole data life cycle, from data management, reproducible analysis, and FAIR sharing. The unit provides access to a computational notebook infrastructure (Jupyter) and a content versioning system.

Head	Data Steward	MSc Trainee	MSc Trainee
			
Tiago Paixão	Carolina Ventura Costa	Mariana Duarte	Rafaela Guerra

Research Data Management @ IGC: data life-cycle

in bold – already in place

Plan	Collection	Store	Use	Share	Archive	Destroy

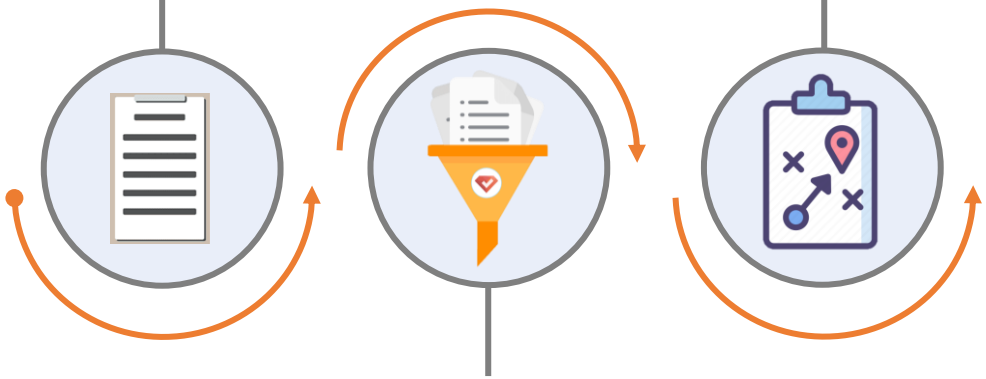
• **Institutional Policies** w/ guidelines on:
  • Sharing & licensing policies
  • Deposition policies
  • Long term storage
  • Curation & annotation policies
	• **Agendo:** platform to book equipments + place facility requests
	• **Electronic Laboratory Book** (mandatory since March 2021)
		• **Centralized Storage**
		• Versioning system
		• Data hub
			• Versioned and reproducible analyses	• **Curation and annotation**
				• Specialized repositories
				• **Institutional Data repositor** (Zenodo)
					• **Long-term storage**	• **10 years**
						• **Transfer to public repositories**

Use case methodology

1
Survey of current practices and issues regarding Research Data Management (RDM)

2
Identified the main problems and requirements

3
Define the action pan



4 - Electronic Laboratory Notebook (eLab)

5 - Research Data Management (RDM)

6 - Informatic infrastructure awareness (ITA)

Survey current **practices, needs and issues** regarding **Research Data Management**

25/27

research groups

43

researchers interviewed

16	Lab managers
8	PIs
7	Postdocs
5	Technicians
4	PhD students
3	MSc students

1. Is there a Data Management Plan within the research lab?
2. Do you follow a naming convention?
3. Do you store metadata along with your data?
4. Where do you store your data? Do you follow
5. Do you keep the raw data?

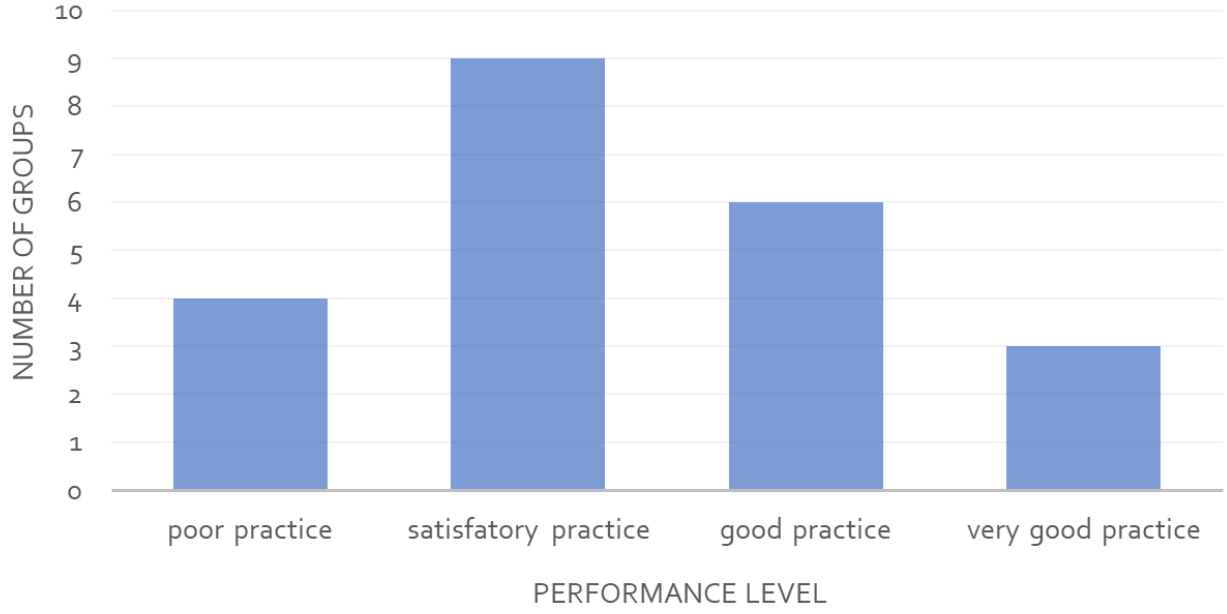
…

What are your suggestions/requirements?

- Strengths – e.g., Good Practices
- Weaknesses – e.g., Knowledge gaps
- Opportunities – e.g., Requirements to develop new features or workflows (space to evolve)
- Threats – e.g., Identified Problems

Research labs performance on **Data Managment**

Research labs performance on **eLab**



Support was provided to different groups on:

Electronic Laboratory Notebook (eLab)

helpdesk support; optimise the use of eLab; create support documentation; testing software and fixe problems

Research Data Management

data storage & management support; folder organization; labelling convention; metadata structure

Informatic infrastructure awareness

filling the gaps between researchers and IT infrastructure

Electronic Laboratory Notebook

- **Documentation page for eLab** with tutorials, templates, guidelines for users and lab managers, FAQs…

available on a **dedicated page** in the intranet

Laboratory Books

Firstly, is important to understand what a lab notebook is and why it is important to keep a good notebook notebook is a complete record of procedures, the reagents you use, the observations you make, and the relevant processes that would enable another scientist to reproduce your results. This includes an explan of why the experiment was performed, how it was performed and the results obtained. Keep in mind that lab notebook is a legal document and can be closely scrutinized in case your research contributes to the of a patent. Also, if any allegations of fraud are brought against your work, your lab notebook can be usec validate your findings. After you have moved on from the lab, your notebook will remain, and is imperative those who comes after you are able to replicate what you have done.

Electronic Laboratory Books (eLab) are mandatory in IGC since March 2021. With an electronic lab bo organise and access your data easily, as well as it has many more benefits compared to traditional paper groups eLab books have already proved to be an excellent tool:

- To increase standardization of procedures (e.g., use templates for standard experiments);
- To distribute new versions of protocols or other relevant documents to every lab member;
- To find researchers´ older data, and;
- To share experiments and protocols with other users using direct links.

As a new member of a research group, you will have access to research data stored, and contribute with the electronic lab notebook adopted at ICG and getting started with it is a very simple and straightforward

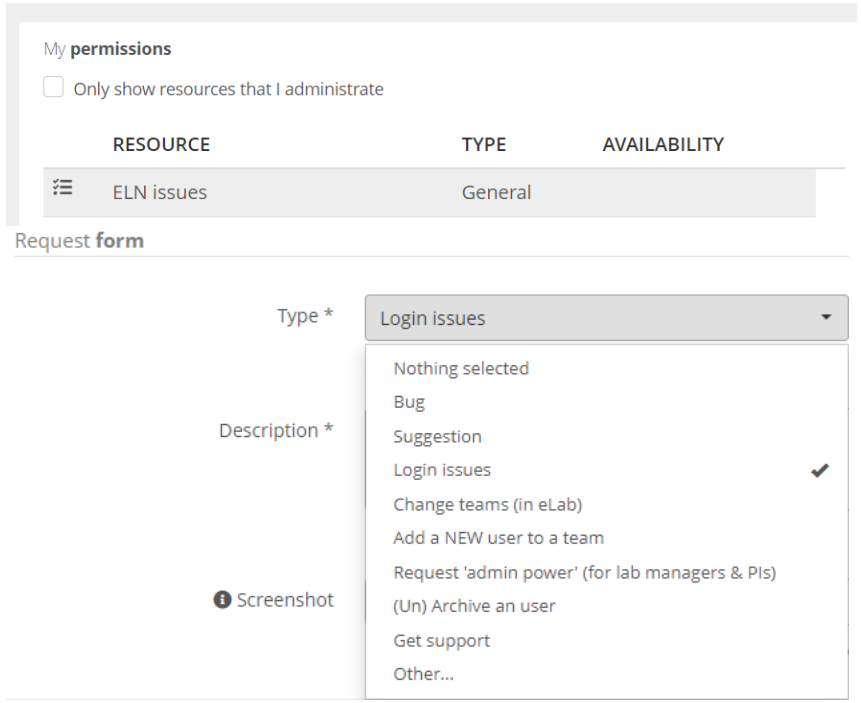
Welcome Session video

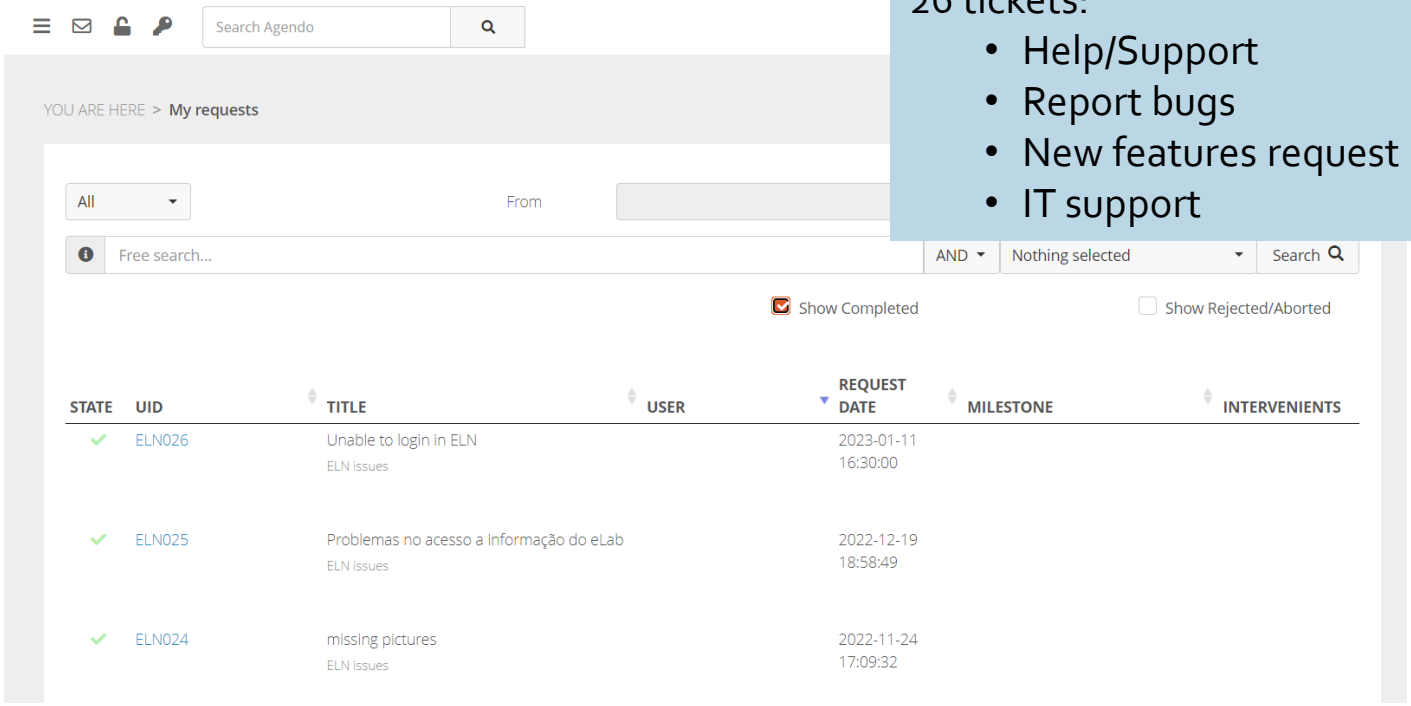
Login procedure

1. Your lab manager has to associate your email account to the research team {*Please find below a detailed description for lab managers*}
2. After that, go to https://labbook.igc.gulbenkian.pt/login.php (*no registration is needed*)
3. Login with the LDAP option, using your **IGC credentials** (*this requires VPN connection if you are not connected to the IGC network*).
4. Start using eLab

Table of Contents

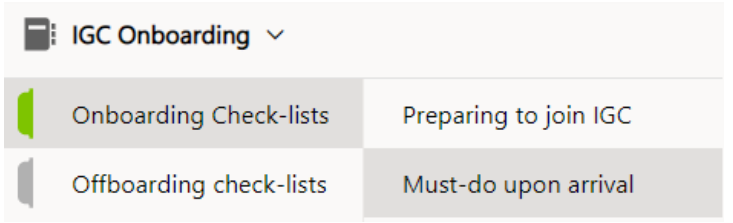
- **Dedicated interface** to report a problem, make a suggestion, ask a question or talking through requirements *(via **Agendo**)*

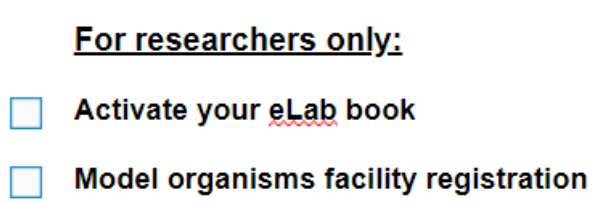




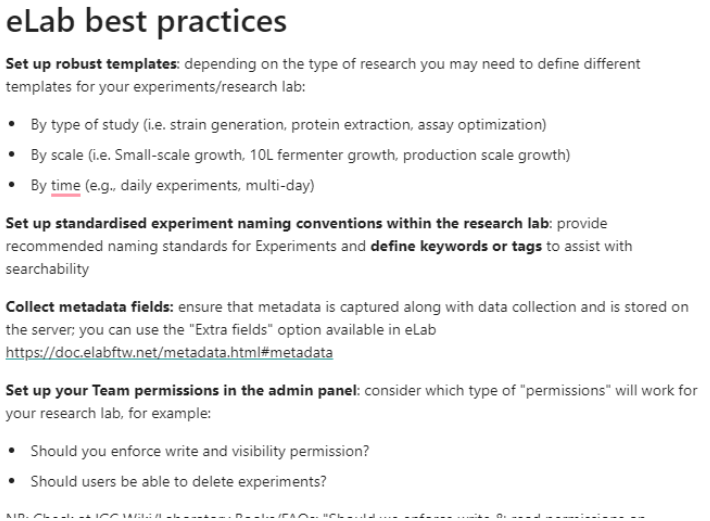
26 tickets:
- Help/Support
- Report bugs
- New features request
- IT support

- OnBoarding training:
  - Welcome Video – overview of eLab
  - Follow-up meeting in collaboration with Human Resources





- eLab committee (w/ lab managers)
  - Regular meetings
  - Responsible for establishing best practices



eLab best practices

Set up robust templates: depending on the type of research you may need to define different templates for your experiments/research lab:

- By type of study (i.e. strain generation, protein extraction, assay optimization)
- By scale (i.e. Small-scale growth, 10L fermenter growth, production scale growth)
- By time (e.g., daily experiments, multi-day)

Set up standardised experiment naming conventions within the research lab: provide recommended naming standards for Experiments and **define keywords or tags** to assist with searchability

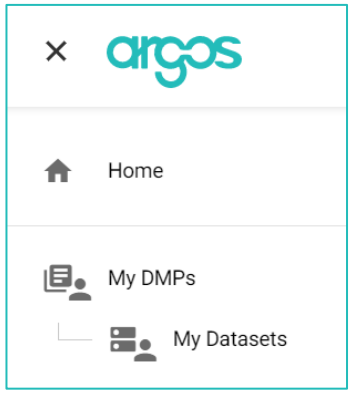
Collect metadata fields: ensure that metadata is captured along with data collection and is stored on the server; you can use the "Extra fields" option available in eLab https://doc.elabftw.net/metadata.html#metadata

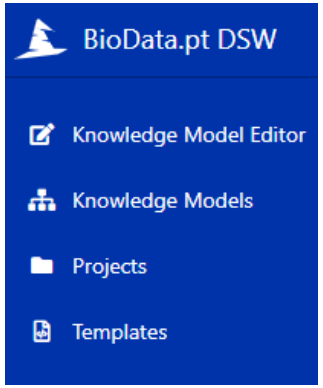
Set up your Team permissions in the admin panel: consider which type of "permissions" will work for your research lab, for example:

- Should you enforce write and visibility permission?
- Should users be able to delete experiments?

Research Data Management

- **Support on DMPs:** in-person support and share available tools to create Data Management Plans





- IGC Data Management Plan Template - in case the funding body does not require a DMP, one shall be provided to IGC regardless.

INSTITUTO GULBENKIAN DE CIÊNCIA

PROJECT NAME – Data Management Plan

Project acronym: ACRONYM

Version number: initial DMP v1

Introduction

Dataset summary

Brief summary of the datasets to be collected

Objectives of the project

Brief summary of the project

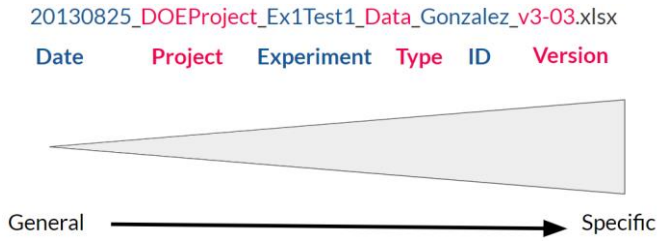
Types and format of datasets

The project will follow institutional guideline on best practices for data formats. Namely, open data

Are you familiar with these feelings?

- I cannot find this file!
- What did I call it again??
- Where is my file??
- What version was it??
- Where is my RAW!!! data???

- **Personalized RDM consulting for research labs:** face-to-face support and customised guidance for folder structure and labelling convention




```
File naming convention:
{a description of the file naming structure adopted}

ProjectAcronym (two capital letters)
TaskNumber (one number from 1 to 9)
------------------------------------------------------------
Date (YYYYMMDD)
User ID (e.g., João Costa = JC)
Strains/Organoids
Subjects: Aplical Nuclear Migration (e.g., ANM)
Condition: Wild-Type (WT) / Control (C)


20220629JC_SXXXX_ANM_WT.xxx
```

```
PROJECT_FOLDER:
/file path/PROJECT ACRONYM
└─ "ProjectLevel_ReadMe" {this document}
└─ Task Number
{this section does not have to be available for users}
------------------------------------------------------------
└─ User folder
      └─ Subject 1: Aplical Nuclear Migration
            ├─ Aquisition Method 1: LigthSheet
                  ├─ Raw Data
                  └─ Processed Data
            ├─ Aquisition Method 2: Genomic Facility
                  ├─ Raw Data
                  └─ Processed Data
            └─ Aquisition Method 3: Other
      └─ Subject 2
      └─ Subject 3
```

Develop the **Data Management** page at IGC Wiki

Before

Book Creator
Add this page to your book Manage book (0 page(s))

Data Management at IGC

IGC is committed to open, rigorous, and reproducible science and therefore the guiding light of Data Management at IGC will be the FAIR (Findable, Accessible, Interoperable, and Reusable) and RRI (Responsible Research and Innovation) principles. IGC considers all publicly funded research data a public good, produced in the public interest, and should be made available in a timely manner, with as few restrictions as possible, but tak property of its researchers.

Grant submission procedures

- make sure you fill out your Data Management Plan
  - Estimate how much storage you will need
  - Point out any ethical or GDPR (personal data) issues with your datasets
  - Define criteria for significance of data
- make sure you account for storage and other Data Management costs in your budget
  - Price is 30 euros/Tb/year
  - accounted as internal services (same as for budgeting an internal facility cost, such as microscopes, v

Resources

- Policy Guidelines
- initial DMP template for grant submissions

After

Research Data Management at the IGC

IGC is committed to open, rigorous, and reproducible science and therefore the guiding light of Data Management at IGC will be the FAIR (Findable, Accessible, Interoperable, and Reusable) and RRI (Responsible Research and Innovation) principles. I research data a public good, produced in the public interest, and should be made available in a timely manner, wi but taking into account the intellectual property of its researchers.

Find here the Policy Guidelines for IGC. A document that describes a set of poli Data Management at the institute.

This page will provide practical guidance on the different data life cycle stages.

Source: RDMkit by ELIXIR-CONVERGE

PLAN

Before you start your project you should define a strategy for managing data and documentation generated within the project. All research performed at IGC shall be governed by its own project-specific DMP, submitted at the time of the grant submission. If the funder does not require it, **one shall be provided to IGC regardless**. For all projects dealing with personal data or human samples, the DMP is a required component of ethical review. Special attention should be given to personal data and compliance to General Data Protection Regulations (GDPR) https://www.gdprsummary.com/. Data Management Plans (DMP) may vary from funder to funder, but they usually share common components, namely:

1. General information about the project
2. Datasets description
3. Ethical and Legal Issues, such as privacy, intellectual property and licences
4. Data Documentation and Metadata
5. Storage solutions, data security and preservation strategy during and after the project
6. Sharing of the data and reuse
7. Costs and resources needed for data management
8. Resposabilities and Resources

(Adapted from RDMkit by ELIXIR-CONVERGE and VIB)

Tools to create your Data Management Plans (DMP)
Multiple DMP tools are available to the community, some are focused exclusively on following funding agencies templates, while others are independent and offer flexibility:

- DMP template for grant submissions at IGC
- Argos online tool https://argos.openaire.eu/
- DS Wizard tool https://researchers.ds-wizard.org/ (Suport information: https://ds-wizard.org/data-management-plans)

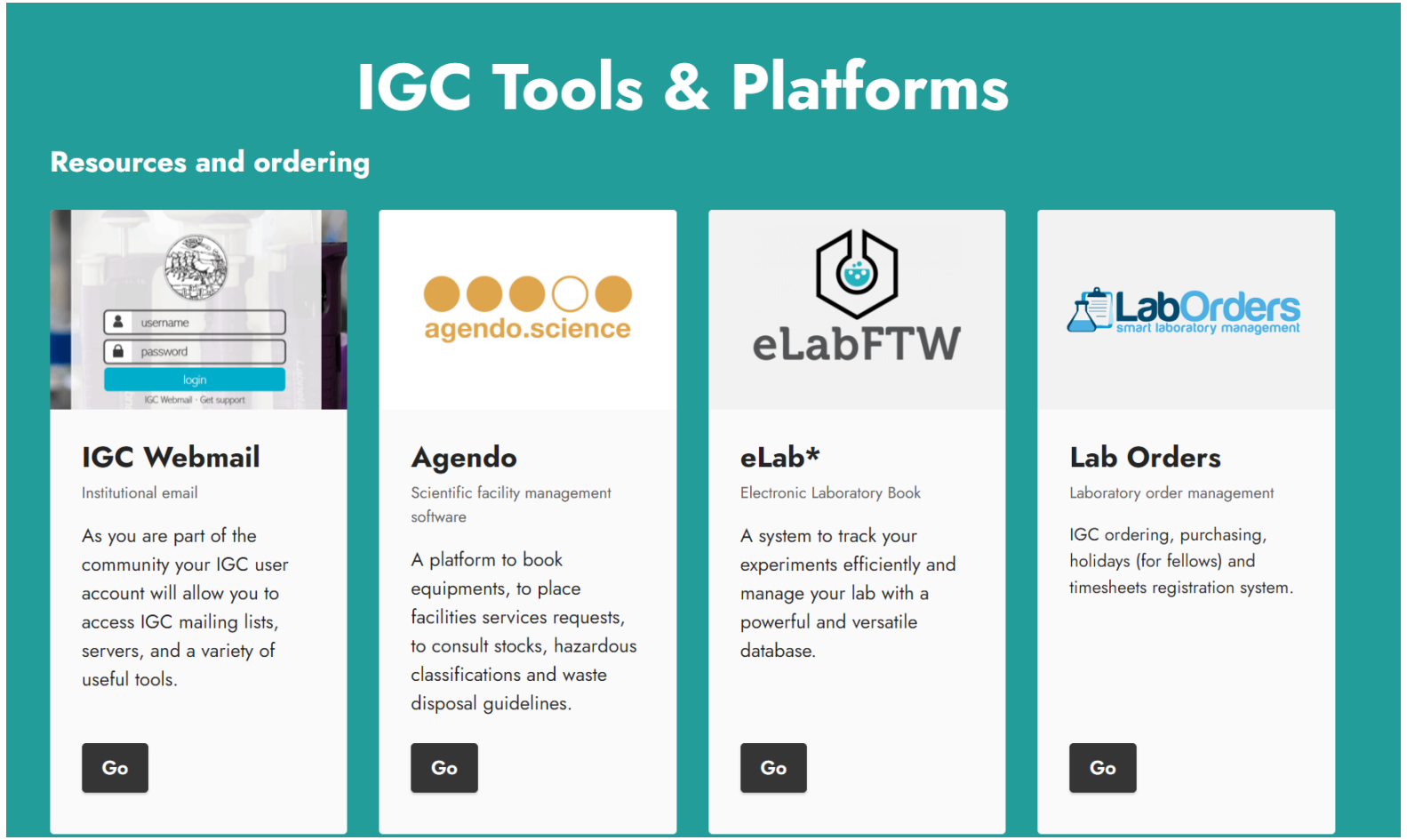
1. General Information

This section is related to practical details concerning your DMP (your funder, project number, name and acronym, PI, relevant legislation under which the funding is granted (national/EU/the funder's country of origin), the requirements of the funder regarding research data management (record keeping, disclosure of results, IP management), etc.

2. Datasets description

- Improve day2day research data management

Centralized page with all tools & platforms available at IGC



IT infrastructure awareness

- Translate IT concepts to researchers, for example the difference between 6-month backup & Storage

IT backups information

eLab snapshots are **daily** performed and stored for a period of **3 months** (in cluster 2). Once a week is performed a synchronisation from cluster 2 to cluster 3 that is kept for 6 months.

*informations provided by IT technicians on 28/06/2022

- Increase awareness among researchers regarding IGC server rules
- eLab database upgrades

DAY 2 DAY DATA MANAGEMENT COURSE
FOR LIFE SCIENCES

Questions?

BioData.pt elixir PORTUGAL

Ready for BioData Management?

DAY 2 DAY DATA MANAGEMENT COURSE
FOR LIFE SCIENCES

Acknowledgements

- ELIXIR Implementation Study "Impact evaluation at Node-level - getting it done"
- ELIXIR-CONVERGE Project “Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services”

BioData.pt elixir PORTUGAL

Ready for BioData Management?

DAY 2 DAY DATA MANAGEMENT COURSE

FOR LIFE SCIENCES

Organization:

Ana Portugal Melo - INESC-ID / BioData.pt

Carolina Costa - IGC / BioData.pt

Daniel Faria - INESC-ID / IST / BioData.pt

Jorge Oliveira - INESC-ID / IST / BioData.pt

@BioData_PT

@BioData.PT

@BioData-PT

www.BioData.pt

inesc id lisboa | GULBENKIAN CIÊNCIA | TÉCNICO LISBOA | BioData.pt

elixir CONVERGE