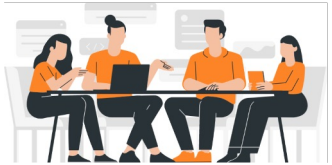
Ready for BioData.pt Management?

Intensive Course

Data Collection

Daniel Faria, Jorge Oliveira, Gil Poiares-Oliveira

BioData.pt | elixir PORTUGAL



I – Challenges

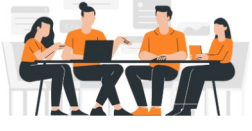
Learning Outcomes:

- Tackle the RDM challenges that arise in data collection

Data Collection

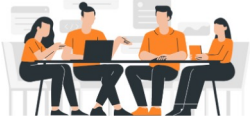
- **Challenges:**
  - Experimental design
  - Data quality
  - Data documentation (provenance)
  - Data organization
  - Data storage
  - Permissions and consent
  - Data protection and security

Ready for BioData.pt Management?



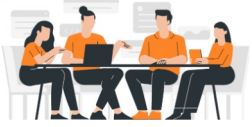
Experimental Design

- Should be mostly defined during the planning stage, but often overflows to the collection stage
  - Adjustments during collection are needed to account for “real life”
  - But we must safeguard the scientific integrity of the experimental design (e.g. biological and technical replicates, clear distinction between trial groups)



Data Quality

- Typically researchers contemplate only **accuracy**
  - E.g. calibrate instruments or estimate error of measurements
- But this is only one of the many **dimensions of data quality**!
  - Most others are also relevant for research data collection



Data Quality

- **Consistency:** ensure agreement between repeated observations or between data sources
  - E.g. the blood type of a patient should be the same across studies
  - E.g. the size of a tree in the field should not decrease over time barring weird circumstances

Data Quality

- **Validity:** ensure data conform to expected format and range
  - E.g. the date of birth of a live patient should not be anterior to 1900 (in principle)
  - E.g. the blood type of a patient must be A, B, AB or O followed by + or -

Data Quality

- **Completeness:** ensure all relevant data is captured for all study subjects
  - E.g. all phenotypic parameters were recorded for all plants in a transcriptomic study
  - E.g. the eCRF was completely filled for all study patients (no missing values)

Data Quality

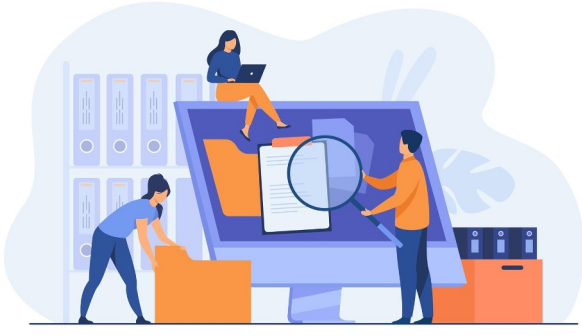
- **Timeliness:** all data is up to date
  - E.g. the clinical data of study patients dates from the moment of sample collection
- **Uniqueness:** data redundancy is minimal or non-existent other than for experimental purposes (reduce storage costs)

Data Quality

- **Integrity:** the raw data is maintained in its original state (no tampering) and any data manipulation is documented and results in a separate file
  - E.g. outlier removal
  - E.g. filling-in missing values through statistical approaches

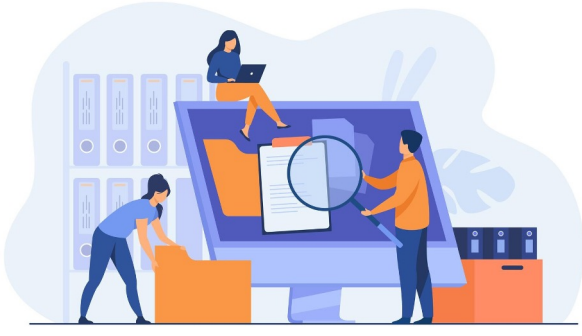
Data Documentation

- At this stage, it is critical to document data provenance
  - E.g. who collected the data, when, and how (experimental metadata)
  - E.g. from what sample/subject is the data derived, what was the sampling procedure, how was the subject selected/treated



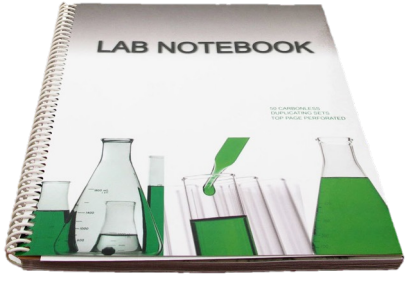
Data Documentation

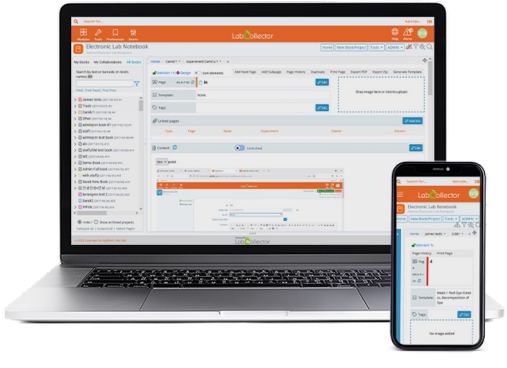
- Favor structured metadata (field-value pairs organized in sections) over long free-text descriptions
- Use ontologies or controlled vocabularies wherever possible
- Link / cross-reference metadata
  - E.g. from the DNA sample to the DNA extraction protocol

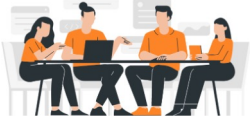


Data Documentation

- Avoid traditional paper documentation in favor of electronic solutions:
  - Electronic Lab Notebooks (ELNs)
  - Laboratory Information Management Systems (LIMS)
  - Electronic Data Capture (EDC) systems







Advantages of Electronic Data Documentation

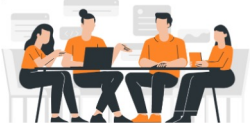
- **Reusability:** you can easily find adapt and reuse previously documented products or procedures
- **Reproducibility:** all the details needed to reproduce an experiment are readily available
- **Sharing & Collaboration:** all group members can have shared access to the documentation environment
- **Oversight:** access can be segmented, enabling PIs to oversee all progress, lab managers to edit protocols and reagents

Ready for BioData.pt Management?

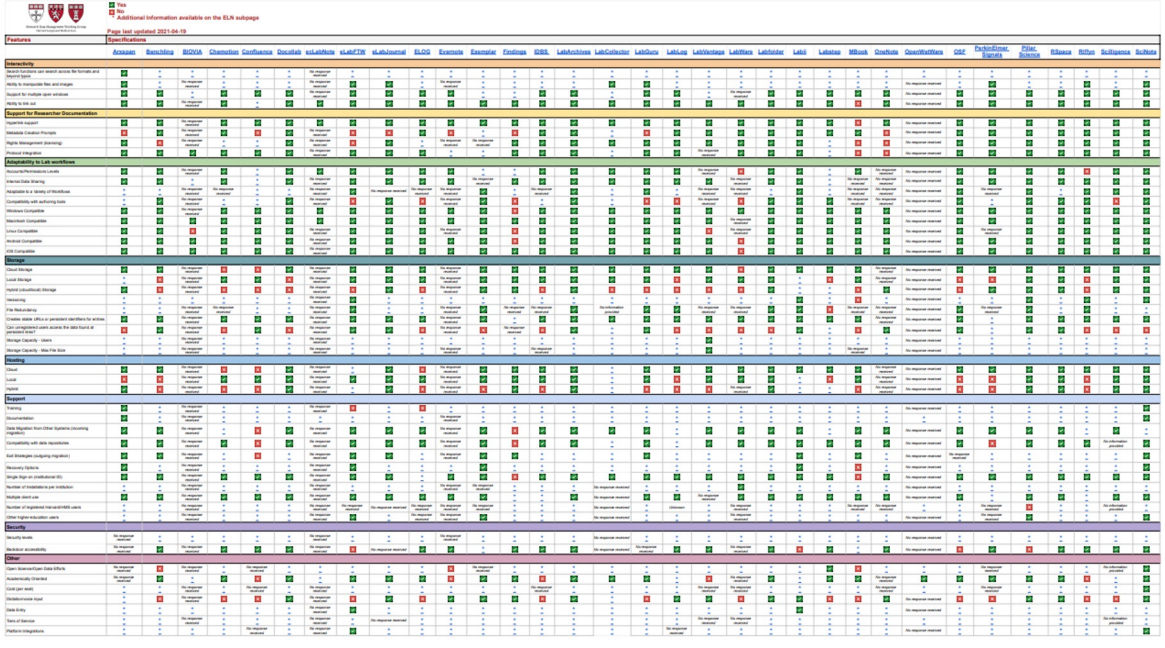
Advantages of Electronic Data Documentation

- **Cross-Referencing:** link from experiments to protocols, samples and results
- **Efficiency:** waste less time writing the same things every day
- **Versioning:** update experiments, protocols, etc, while keeping track of the updates
- **Organisation:** easily keep track of data files; in EDC systems, define the structure of the data
- **Integration:** with institutional facilities and data repositories, data publication platforms

Ready for BioData.pt Management?



Choosing an ELN



Yes
No
Additional information available on the ELN subpage

Page last updated 2021-04-19

Features | Specifications

Acaspan | Benchling | BIOVIA | Chemotion | Confluence | DocuLab | eLabNext | eLabFTW | eLabJournal | ELOG | Evernote | Exemplar | Findings | IDBS | LabArchives | LabCollector | LabGuru | LabLog | LabVantage | LabWare | LabFolder | Labii | Labstep | MBook | OneNote | OpenWetWare | OSF | PerkinElmer Signals | Pilar Science | RSpace | Riffyn | Scilligence | SciNote

Interactivity
Support for Researcher Documentation
Adaptability to Lab workflows
Storage
Hosting
Support
Security
Other

DOI 10.5281/zenodo.4723753

eLabFTW

Benchling

<LAB FOLDER>

Labguru

Confluence

Evernote

Ready for BioData.pt Management?

Choosing an ELN

- **Considerations:**
  - Free vs. Paid
  - Self-Hostable vs. Centralized
  - Open Source vs. Private
  - Easy to Customise and Integrate
  - Support for Metadata Standards & Ontologies
  - Type of Storage Supported

Ready for BioData.pt Management?

Data Organisation

- **Unique Persistent Identifiers:** being able to uniquely identify experimental samples, reagents and procedures, or (meta)data records, variables and values is critical for reproducibility and reusability
  - For internal objects, an institutionally unique identification schema should be defined
  - For domain concepts or published objects, a globally unique identifier should be used (e.g. ontology identifier)

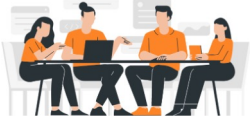
Data Organisation

- **File Naming:** opt for brief but descriptive file names, including some provenance information on name
- **File Versioning:** use a version control system (e.g. git) instead of keeping manual track of file versions
- **File Format:** favor non-proprietary well-established file formats whenever possible
- **Folder Structure:** subfolders should match the experimental workflow and have short unique self-explanatory names; master folder should have a README file detailing the folder structure

Ready for BioData.pt Management?

Data Storage

- We will discuss Data Storage in more depth at the Data Preservation stage
- During the data collection stage, we typically require storage that is:
  - Private and secure (data is for internal consumption only and is usually very valuable)
  - Quick and easy access
  - Easy to connected to computing (for the next two stages of the lifecycle)

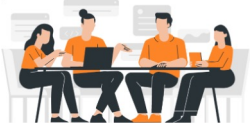


Permissions & Consent

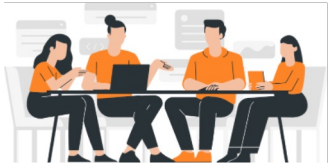
- If your study involves human subjects, the data collection stage is when you must ensure you are actually allowed to collect the data
  - Get approval of the ethics committee of your institution
  - Get informed consent from all participants in your study
  - If you are reusing existing private datasets, get permission to access them

Data Protection & Security

- If your study involves personal data, you must ensure GDPR compliance
  - Even consenting data subjects can ask you to remove their data
  - You must keep all data private and secure until it is anonymised, and personally identifying data private and secure at all times
- If your study involves sensitive data, you must have stringent security measures to avoid data breaches
  - E.g. encrypt the data while it is not being used



Ready for BioData.pt Management?



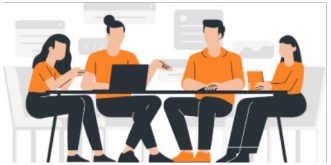
Intensive Course

Data Collection

Thank You!

Questions?

BioData.pt | elixir PORTUGAL



II – Hands-On

Learning Outcomes:

- Use an electronic lab notebook

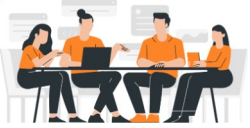
eLabFTW

- The ELN we'll be using for the exercise
- It is free, open source, self-hostable, and easy to use
- It was the ELN adopted institute-wide by the IGC
- We have an instance available for training at elab.biodata.pt
  - It is periodically cleared, so do not use it beyond this course

Mock Project

All hands-on exercises for this course will be based on the mock project available here: biodata.pt/mock

Please read through the project before proceeding!



Group Exercise

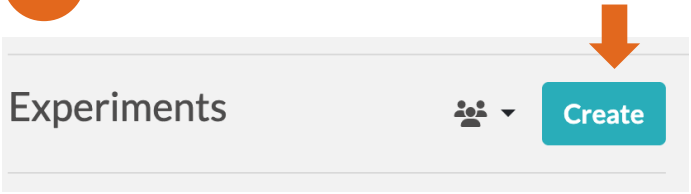
Document the whole genome sequencing task from the mock project (WP1, Task 2) in our training ELN, from patient samples to DNA samples, to sequencing results

1. Navigate to elab.biodata.pt
2. Create an account and log in
3. Start documenting!

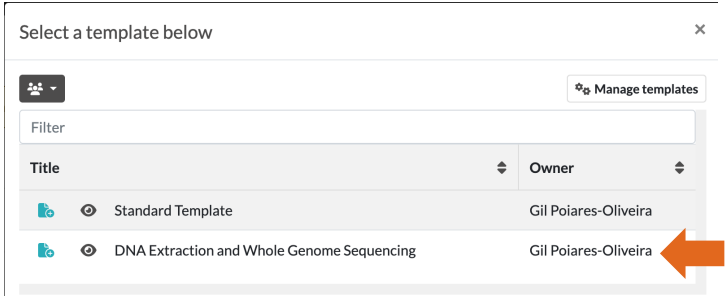
Ready for BioData.pt Management?

Possible solution

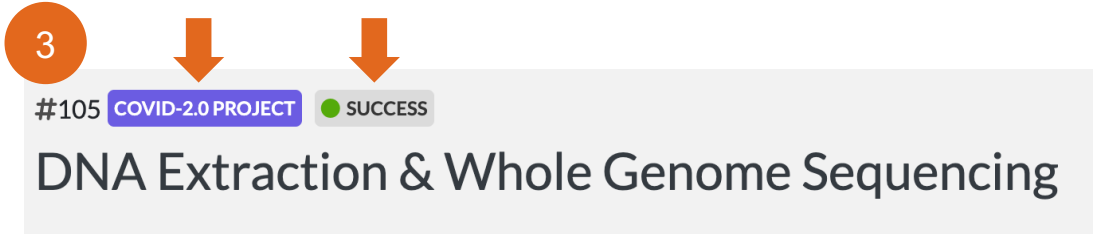
1



2



3



4

5

Make it only visible to project colleagues

Ready for BioData.pt Management?

Possible solution

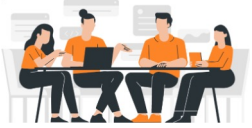
Objectives:

This protocol aims to perform full genomic DNA sequencing from patient buccal samples.

Samples:

Samples were provided by Hospital X and Hospital Y from COVID-2.0 patient buccal swab samples, labelled X-001~X-200 and Y-001~Y-200, respectively. Sample identification table can be found at `patients-1.tre.infra.biodata.pt` on path `/data/projects/covid-2/samples/sample_ids.csv`.

If sample information is not in eLab, indicate where to find it



Possible solution

Procedure:

Day 1: 2024-01-22 ← Separate by day

Genomic DNA extraction was performed according to standard protocol. ← Link to protocol

Samples were quantified at the Nanodrop.

Samples were stored at -20°C. ← Link to storage location

Link to reagent →

Day 2: 2024-01-23

Library prep for Illumina sequencing was performed as per Illumina DNA PCR-Free Library Prep protocol, using the corresponding prep kit.

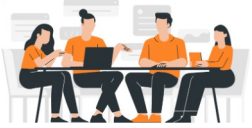
Sequencing was performed using Illumina NextSeq 2000. ← Link to equipment

Resulting FASTQ files were encrypted on site using following command:

```
gpg --encrypt --recipient human-sequencing@biodata.pt covid-2_samples.tar.gz
```

← bash code sample

Files were then uploaded to sequencing server using encrypted FTP.



Possible solution

Observations/Results:

- DNA quantification at Nanodrop yielded sufficient amounts of DNA per sample (results attached). Some samples, however, had considerably lower concentrations, although sufficient for the purposes.
- Library preparation went as expected, according to manufacturer guidelines.

Datasets:

Resulting FASTQ files can be found at `sequencing-1.tre.infra.biodata.pt` at path `/data/sequencing/covid-2/SEQ-COVID2-001/encrypted-samples`.

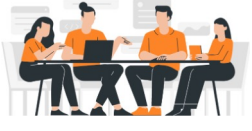
Results dataset location for large files

ATTACHED FILE

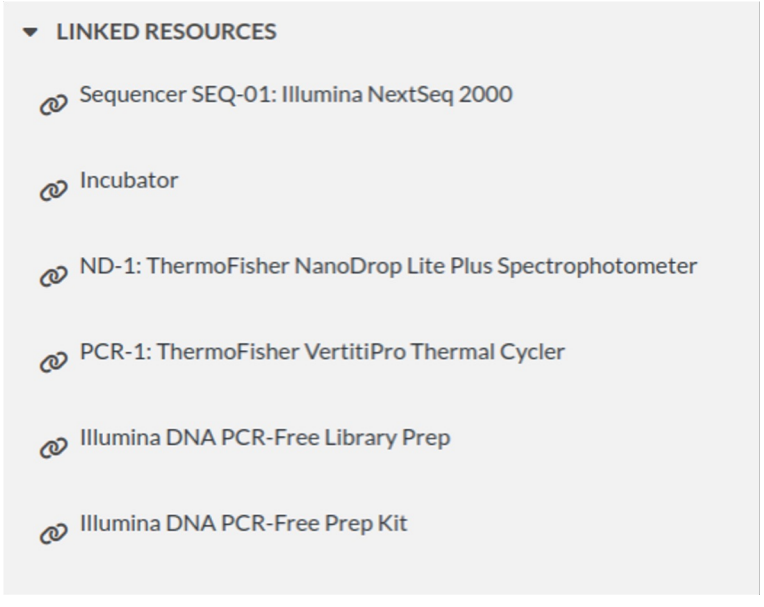
nanodrop_results_0012213.csv 43.00 B - 2024-01-27 10:55:58

DNA concentrations measured with Nanodrop

Ready for BioData.pt Management?



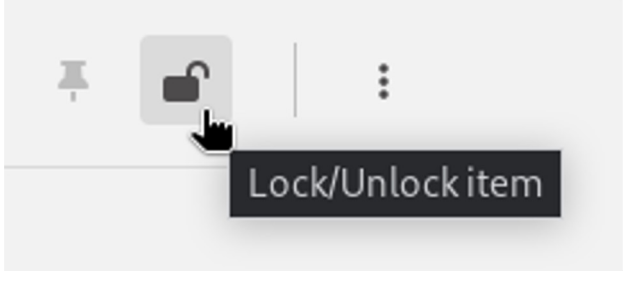
Possible solution



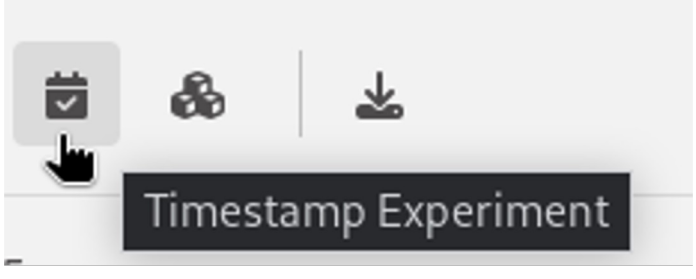
Linked resources (equipment/reagents/protocols) used in this experiment

Let's make it official

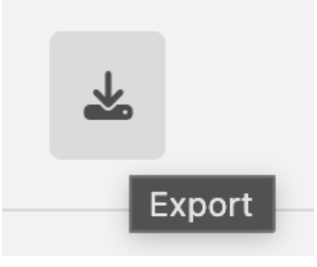
Finalise your experiment by locking it and adding a timestamp that proves its state at the time



1. Lock it



2. Timestamp it

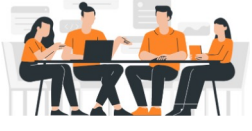


3. Download the proof

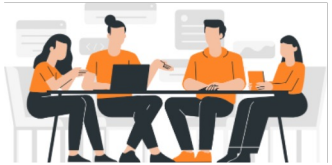
eLabFTW reaches out to a trusted third party to save a snapshot of your experiment at the time of timestamping to avoid fraud

Possible solution

Get the PDF here
biodata.pt/elab-results



Ready for BioData.pt Management?



Intensive Course

Data Collection

Thank You!

Questions?

BioData.pt | elixir PORTUGAL