Ready for BioData.pt Management?

Intensive Course

Data Preservation

Daniel Faria, Jorge Oliveira, Gil Poiares-Oliveira

BioData.pt | elixir PORTUGAL

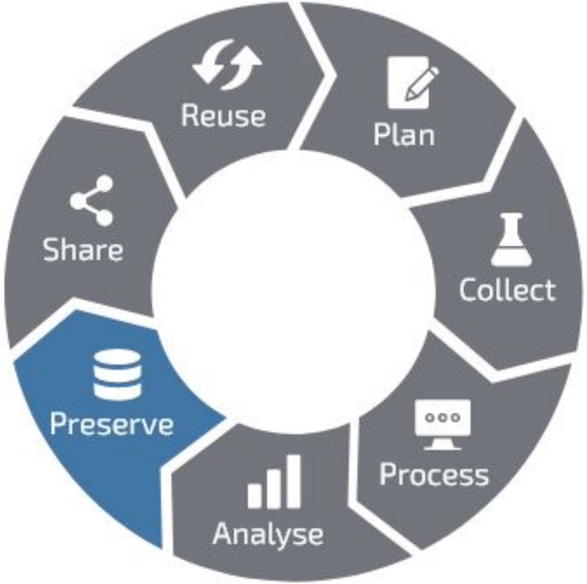
I – Challenges

Learning Outcomes:

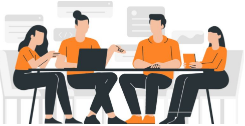
- Tackle the RDM challenges that arise in data preservation

Data Preservation

- The stage at which the only concern is long-term storage of relevant data (and eventual disposal of irrelevant data)
- It overlaps with **Data Sharing:** we can preserve data by depositing it in a public repository
  - In this chapter, we'll focus only on the local preservation scenario

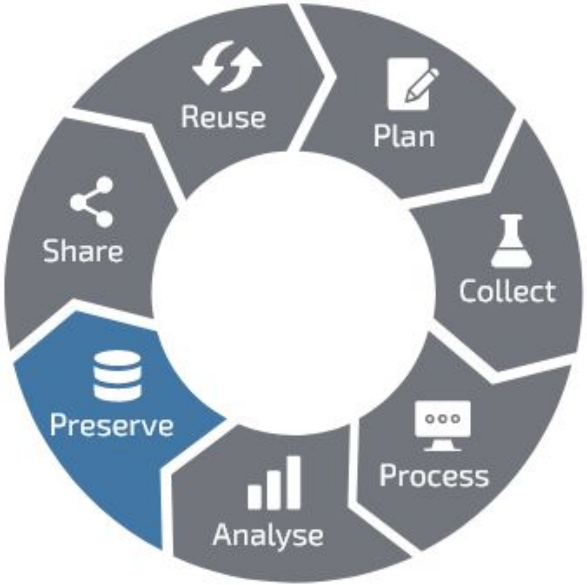


Ready for BioData.pt Management?



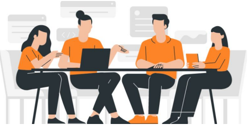
Data Preservation

- **Challenges:**
  - Requirements & Relevance
  - Volume, Duration & Accessibility
  - Security
  - Documentation & organization



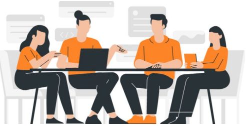
Requirements & Relevance

- To identify what data to preserve (and for how long) you must consider:
  - **Requirements** for data preservation:
    - Legal or ethical (e.g. data from clinical trials)
    - Funders' (typically require data preservation for 5 or 10 years after the end of the project)
    - Institutional



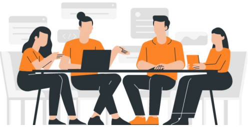
Requirements & Relevance

- To identify what data to preserve (and for how long) you must consider:
  - The **relevance** of the data:
    - Value to society (scientifically, historically or culturally significant data)
    - Uniqueness and difficulty to re-generate
    - Potential for reuse



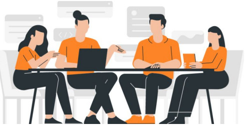
Requirements & Relevance

- Raw data are usually more relevant
- Intermediate data processing and analysis files can often be discarded, if they can be easily reproduced from the raw data (if the data processing and analysis workflow is preserved)
- Final result files may also be relevant
- Temporary or mutable data should generally not be selected for preservation



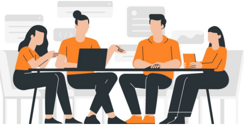
Volume, Duration & Accessibility

- The three main factors that determine storage costs:
  - **Volume** is dictated by the data and directly affects the required storage capacity
  - **Duration** is usually dictated by external requirements and is a factor due to the limited lifetime of storage media as well as the energy costs
  - **Accessibility** is dictated by the expected need to access the data in the future, which affects the choice of storage media



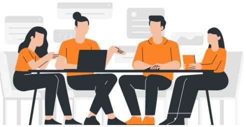
Volume, Duration & Accessibility

- Volume can be reduced by compressing the data, at the cost of accessibility
- Duration is less expensive if expected access is low:
  - HDD and SSD disks last longer
  - Tape becomes an option and is cheaper than disks
- Under accessibility we must also consider:
  - Expectancy of changes to the data (low at this stage)
  - Acceptable recovery time in case of disk failure



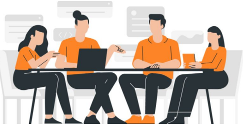
Security

- The fourth factor affecting storage costs
- “Standard” security includes disk redundancy and backups on tape
- If data is sensitive, data encryption is recommended as additional security
- If access to the data is restricted, accessibility costs will likely be higher



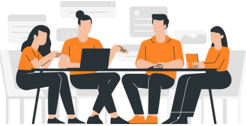
Documentation & Organization

- Data selected for preservation requires particular care with respect to both documentation and organisation
    - At the very least, the IT staff need to be able to identify the data and data owner
    - In case of an audit, the auditor needs to be able to identify and understand the data
    - The data owner needs to be able to identify and understand the data after a few years have passed



Documentation & Organization

- If we did a good job documenting and organizing the data during the previous stages, then little or no work is needed
- Include a README file at the head of the file structure describing the dataset and the file structure
  - Include links to protocols, workflows, and all relevant external references in the README file
  - Alternatively, protocols and workflows can be included in the file structure



Thank You!

Questions?

II – Hands-On

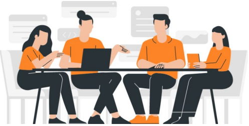
Learning Outcomes:

- Define data storage requirements
- Estimate data storage costs

Data Stewardship Wizard (DSW)

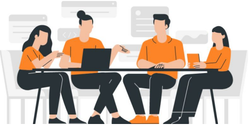
DSW Storage Costs Evaluator

- A web tool for data management planning developed by DTL-NL (which we'll explore in more detail later)
- Includes a service to estimate storage costs: https://storage-costs-evaluator.ds-wizard.org/
- You can also get the source code from https://github.com/ds-wizard/storage-costs-evaluator



Group Exercise

- Open the DSW Storage Costs Evaluator at https://storage-costs-evaluator.ds-wizard.org/
- Use the tool to estimate the cost of data preservation for the mock project
  - Note that different datasets may have different storage needs, namely with respect to accessibility and security

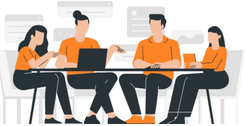


Solution 1

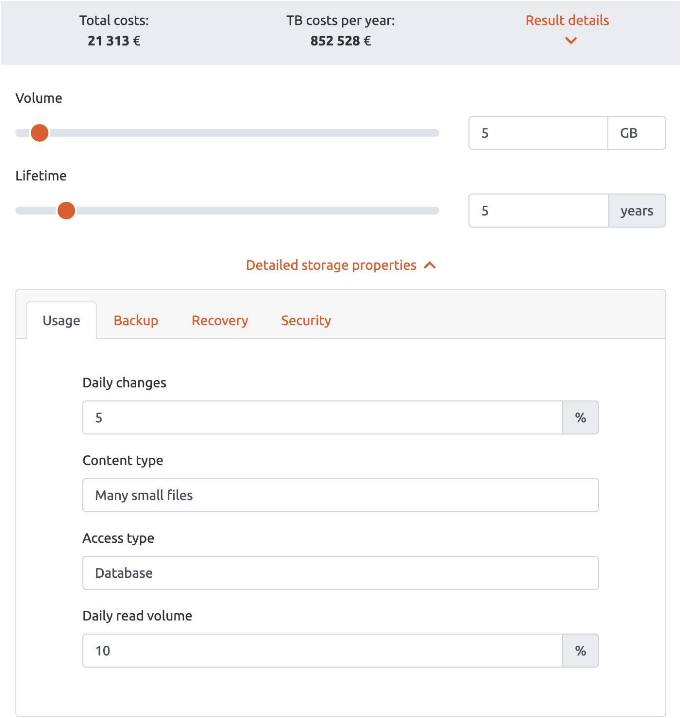
1st - decide preservation time: assumed 5 years

Type of Data	Size	Cost
Clinical data and metadata	< 5 GB	21 313 €
RAW fastq	~40 TB	28 177 € or for free*
Analysed	~300 GB	26 957 €

*if deposited in public repository as it is created



Clinical Data & Metadata



Total costs: **21 313 €**

TB costs per year: **852 528 €**

Volume: 5 GB

Lifetime: 5 years

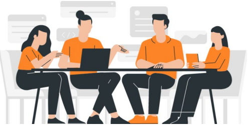
Daily changes: 5 %

Content type: Many small files

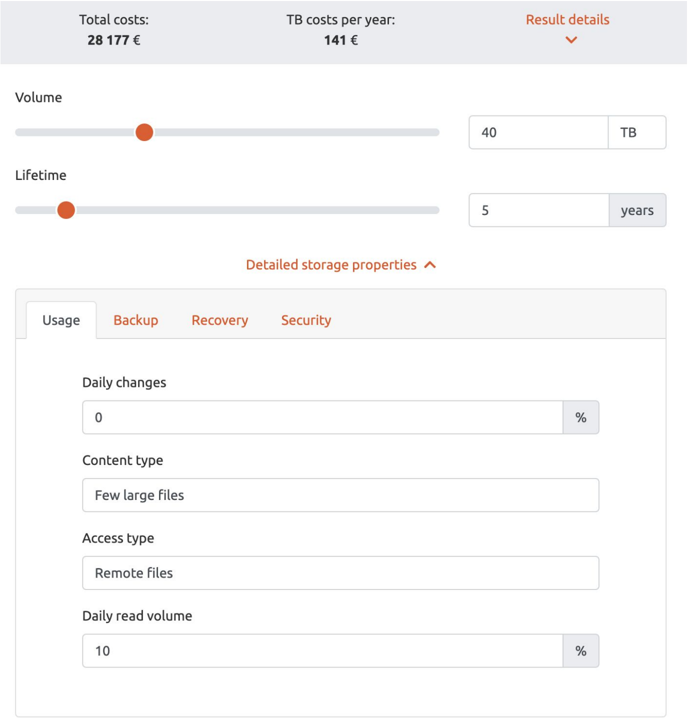
Access type: Database

Daily read volume: 10 %

- Minor occasional changes to data
- Tape backup w/ low frequency
- Flexible recovery time
- Privacy sensitive
- Authorized people access



RAW Data



Total costs: **28 177 €**

TB costs per year: **141 €**

Volume: 40 TB

Lifetime: 5 years

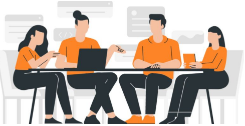
Daily changes: 0 %

Content type: Few large files

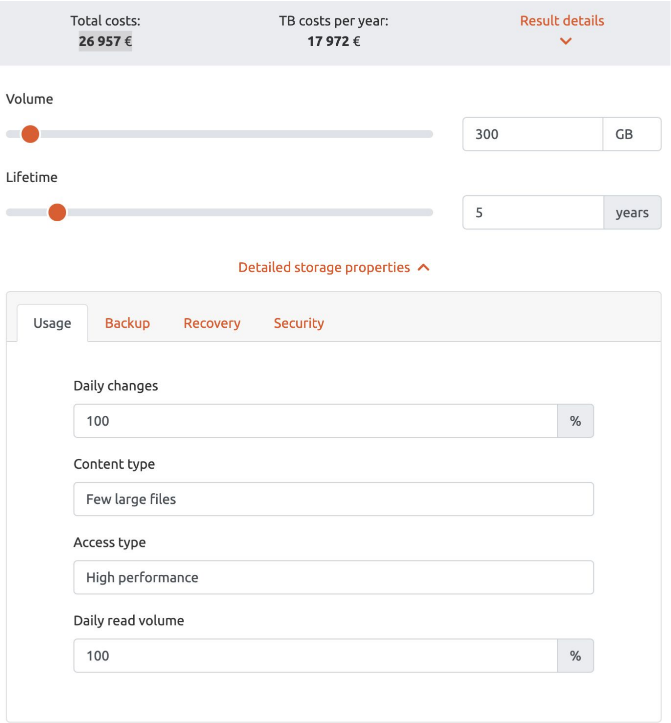
Access type: Remote files

Daily read volume: 10 %

- ○ Few or no changes to data
- ○ Tape backup w/ low frequency
- ○ Flexible recovery time
- ○ Privacy sensitive
- ○ Authorized people access



Analysis Data



Total costs: **26 957 €**

TB costs per year: **17 972 €**

Volume: 300 GB

Lifetime: 5 years

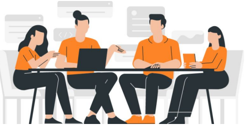
Daily changes: 100 %

Content type: Few large files

Access type: High performance

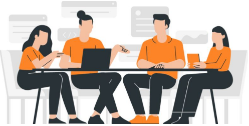
Daily read volume: 100 %

- ○ Many changes / read of data
- ○ Backup w/ high frequency
- ○ High performance
- ○ Privacy sensitive
- ○ Authorized people access



Solution 2?

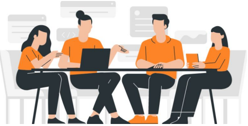
Can we have all these datasets in the same infrastructure?



Solution 2

It might be possible to create different storage solutions inside the same infrastructure and save some money:

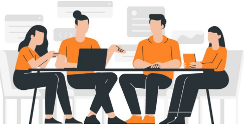
Type of Data	Size	Cost
All project data	~ 41 TB	57 044 €



Solution 3

If we decide to submit all sequencing data to a public repository:

Type of Data	Size	Cost
All project data	< 500 GB	27 160 €



Thank You!

Questions?