BioData.pt

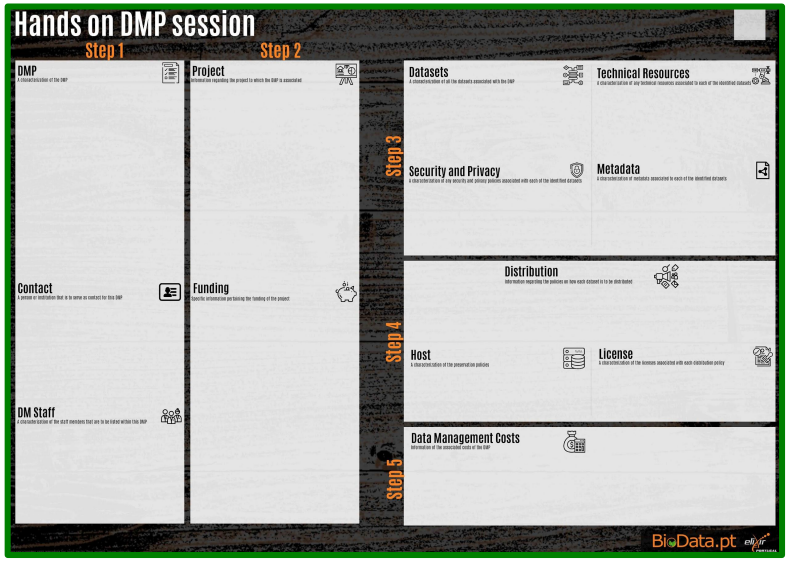
Ready for BioData Management?

Hands-On DMP Exercise

João Cardoso, Daniel Faria

Hands-On DMP Exercise

- The goal of this group exercise is for each group to **create their own DMP** for the provided **mock project.**
- Participants should follow the **DMP Creation Methodology** detailed in the following slides.
- The DMP is to be prepared by collecting the required Information in **post-its** and then posting them onto the corresponding section of the **DMP Canvas**.
- We will provide **help** throughout the exercise in exchange for a **“help token”**.



Hands-On DMP Exercise

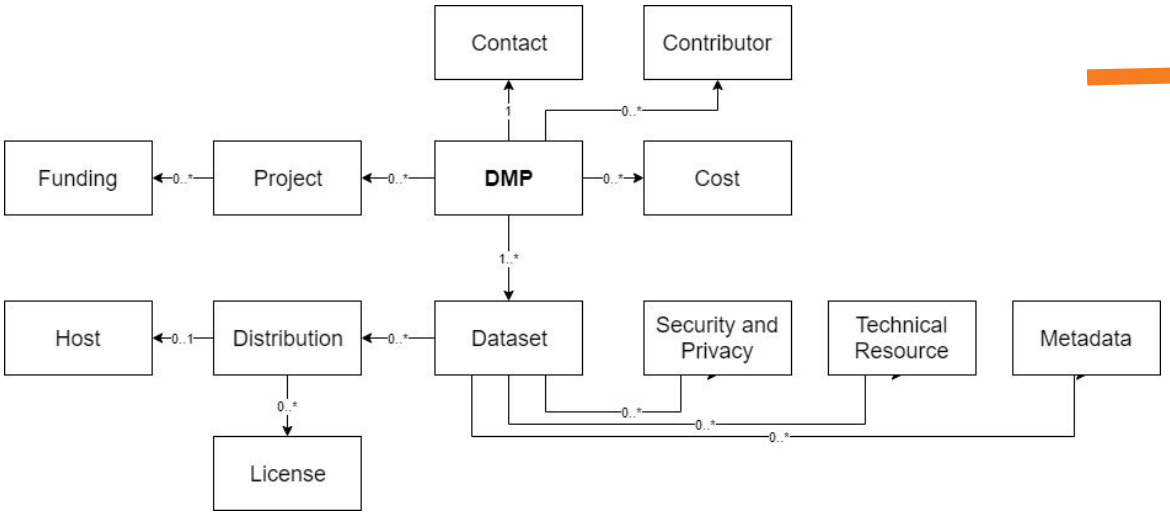
- Keep in mind that:
  - **DMPs are living documents**, information is always **subject to be changed** throughout the process.
  - Do **not feel trapped** by **previous decisions**, and do not be afraid to revise them.
  - **Not all** information is **explicitly described** in the project, you may have to **deduce, look up or make up information**.

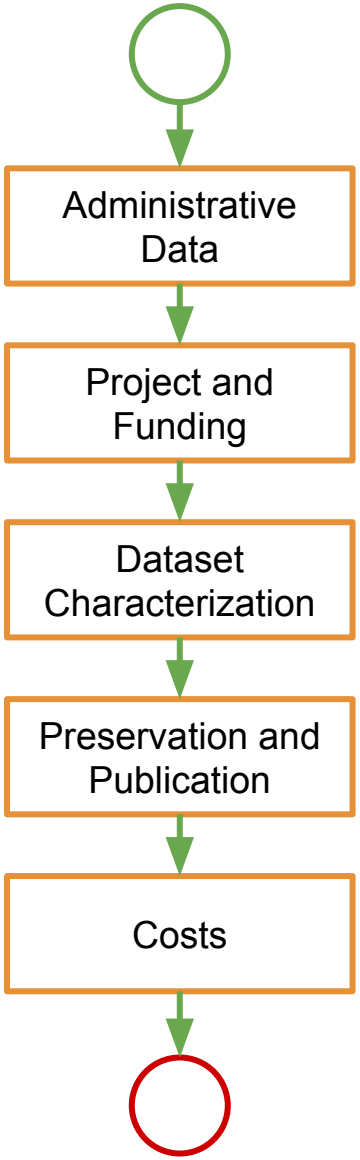
Project X

- **Title:** Unveiling the mechanisms of Disease X
- **Context:** BioData.pt is applying for funding from the FCT.
- **Motivation:**
  - The cause of **Disease X** has been recently discovered to be a **virus**, phage X.
  - It **infects** normal **gut bacteria** and leads them to become virulent and cause **chronic intestinal infection**.
  - This disease has been **spreading rapidly** in Europe, with **costs** in health-care reaching the **tens of millions of Euros**.

Creating a DMP

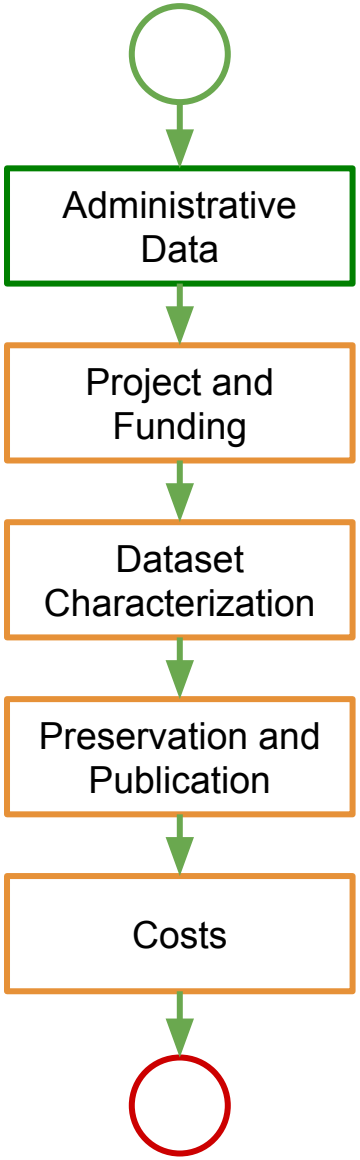
- The **DMP Creation Methodology** comprises 5 steps.
- Each step focuses on a **specific aspect** of the DMP.
- It is **based** on the **RDA's DMP Common Standards** metadata application standard.





Creating a DMP (Step 1)

- **Step 1 - Administrative Data**
  - **Characterization** of the **DMP document**, and the **responsibilities** of all the **people** mentioned.
  - The information is split in three sections:
    - **General information** characterizing the **DMP document**.
    - **Contact** (person or institution) for the DMP.
    - A listing of all **collaborators** and their **roles** in the DMP.
- **No pitfalls here, this section is essentially bureaucratic**



Creating a DMP (Step 1)

In the project

Project X (Application to Fundação para a Ciência e Tecnologia)

Title of the project: Unveiling the mechanisms of Disease X

Participants:

- Prof. Coor Dinator (coor.dinator@biodata.pt) [PI & DMP Coordinator]
- Dr. Dat Manger (dat.manger@biodata.pt) [Data Manager]
- Dr. Col Hector (col.hector@biodata.pt) [Clinical Data & Sample Collector]
- Dr. R. Sercher (r.sercher@biodata.pt) [Researcher]
- Mrs. A. D'Min (a.dmin@biodata.pt) [Project Manager]

Host Institution: BioData.pt

Start date: January 1st, 2021

Duration: 36 months

Creating a DMP (Step 1)

In the project

Project X (Application to Fundação para a Ciência e Tecnologia)

Title of the project: Unveiling the mechanisms of Disease X

Participants:

- Prof. Coor Dinator (coor.dinator@biodata.pt) [PI & DMP Coordinator]
- Dr. Dat Manger (dat.manger@biodata.pt) [Data Manager]
- Dr. Col Hector (col.hector@biodata.pt) [Clinical Data & Sample Collector]
- Dr. R. Sercher (r.sercher@biodata.pt) [Researcher]
- Mrs. A. D'Min (a.dmin@biodata.pt) [Project Manager]

Host Institution: BioData.pt

Start date: January 1st, 2021

Duration: 36 months

- Generic information on the DMP document:
  - Title, institution, start date, duration, etc.
  - Description from the abstract.

Creating a DMP (Step 1)

In the project

Abstract:

The cause of Disease X has been recently discovered to be a virus, phage X, which infects normal gut bacteria and leads them to become virulent and cause chronic intestinal infection. Although non-fatal, this disease has been spreading rapidly in Europe, with costs in health-care reaching the tens of millions of Euros.

This project aims to uncover the mechanisms of disease X by sequencing phage X and studying the effects of its infection in human gut microbiota at the population and molecular level. We will assess which bacterial taxa are infected by phage X and what effect the infection has on the relative abundance of the various taxa, as well as what e
infected taxa at the gene expression level.

The project will be a key step towards improving
potentially being able to cure it.

- Generic information on the DMP document:
  - Description from the abstract.

Creating a DMP (Step 1)

In the project

Project X (Application to Fundação para a Ciência e Tecnologia)

Title of the project: Unveiling the mechanisms of Disease X

Participants:

- Prof. Coor Dinator (coor.dinator@biodata.pt) [PI & DMP Coordinator]
- Dr. Dat Manger (dat.manger@biodata.pt) [Data Manager]
- Dr. Col Hector (col.hector@biodata.pt) [Clinical Data & Sample Collector]
- Dr. R. Sercher (r.sercher@biodata.pt) [Researcher]
- Mrs. A. D'Min (a.dmin@biodata.pt) [Project Manager]

Host Institution: BioData.pt

Start date: January 1st, 2021

Duration: 36 months

- **Contact** (person of institution) for the DMP.

Creating a DMP (Step 1)

In the project

Project X (Application to Fundação para a Ciência e Tecnologia)

Title of the project: Unveiling the mechanisms of Disease X

Participants:

- Prof. Coor Dinator (coor.dinator@biodata.pt) [PI & DMP Coordinator]
- Dr. Dat Manger (dat.manger@biodata.pt) [Data Manager]
- Dr. Col Hector (col.hector@biodata.pt) [Clinical Data & Sample Collector]
- Dr. R. Sercher (r.sercher@biodata.pt) [Researcher]
- Mrs. A. D'Min (a.dmin@biodata.pt) [Project Manager]

Host Institution: BioData.pt

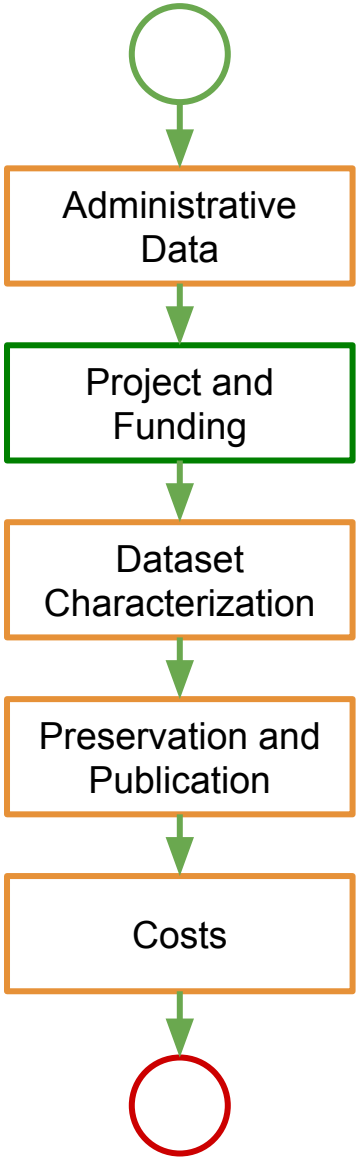
Start date: January 1st, 2021

Duration: 36 months

- A listing of all **collaborators** and their **roles** in the DMP.

Creating a DMP (Step 2)

- **Step 2 - Project and Funding**
  - **Characterization** of the **project(s)** and their sources of **funding**.
  - The information is split in two sections:
    - Information regarding the **project(s)** to which the **DMP is associated**.
    - Information pertaining to the **funding** of a **particular project**.
- **Also no pitfalls here, and again a mainly bureaucratic section**



Creating a DMP (Step 2)

In the project

Project X (Application to Fundação para a Ciência e Tecnologia)

Title of the project: Unveiling the mechanisms of Disease X

Participants:

- Prof. Coor Dinator (coor.dinator@biodata.pt) [PI & DMP Coordinator]
- Dr. Dat Manger (dat.manger@biodata.pt) [Data Manager]
- Dr. Col Hector (col.hector@biodata.pt) [Clinical Data & Sample Collector]
- Dr. R. Sercher (r.sercher@biodata.pt) [Researcher]
- Mrs. A. D'Min (a.dmin@biodata.pt) [Project Manager]

Host Institution: BioData.pt

Start date: January 1st, 2021

Duration: 36 months

Creating a DMP (Step 2)

In the project

Project X (Application to Fundação para a Ciência e Tecnologia)

Title of the project: Unveiling the mechanisms of Disease X

Participants:

- Prof. Coor Dinator (coor.dinator@biodata.pt) [PI & DMP Coordinator]
- Dr. Dat Manger (dat.manger@biodata.pt) [Data Manager]
- Dr. Col Hector (col.hector@biodata.pt) [Clinical Data & Sample Collector]
- Dr. R. Sercher (r.sercher@biodata.pt) [Researcher]
- Mrs. A. D'Min (a.dmin@biodata.pt) [Project Manager]

Host Institution: BioData.pt

Start date: January 1st, 2021

Duration: 36 months

- Information regarding the **project(s)** to which the **DMP is associated**.

Creating a DMP (Step 2)

In the project

Project X (Application to Fundação para a Ciência e Tecnologia)

Title of the project: Unveiling the mechanisms of Disease X

Participants:

- Prof. Coor Dinator (coor.dinator@biodata.pt) [PI & DMP Coordinator]
- Dr. Dat Manger (dat.manger@biodata.pt) [Data Manager]
- Dr. Col Hector (col.hector@biodata.pt) [Clinical Data & Sample Collector]
- Dr. R. Sercher (r.sercher@biodata.pt) [Researcher]
- Mrs. A. D'Min (a.dmin@biodata.pt) [Project Manager]

Host Institution: BioData.pt

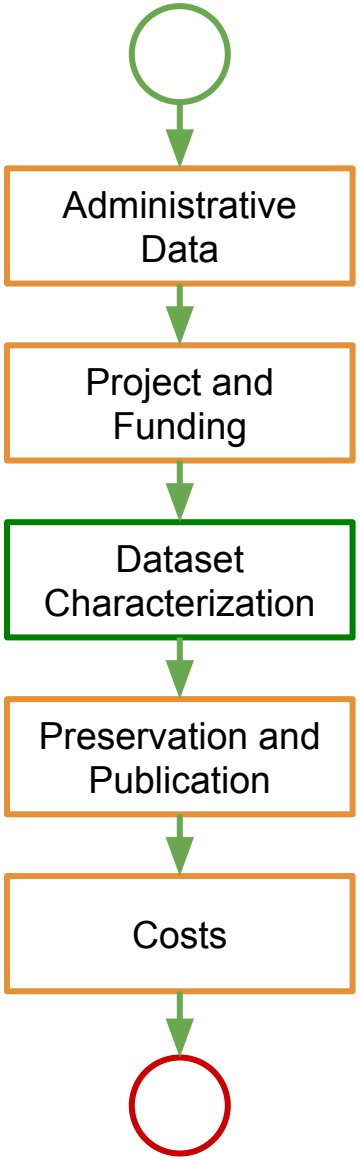
Start date: January 1st, 2021

Duration: 36 months

- Information pertaining the **funding** of a **particular project**.

Creating a DMP (Step 3)

- **Step 3 - Dataset Characterization**
  - **Characterization** of the **dataset(s)** that are encompassed by the DMP. Apart from **generic information** on the dataset, **additional** descriptions of **security and privacy** policies, **technical resources** and **metadata** standards should also be given.
  - The information is split in four sections:
    - **General information** about all **datasets**.
    - Any **security and privacy** policies associated with the datasets.
    - **Technical resources** associated with the datasets.
    - **Metadata** associated with the datasets.



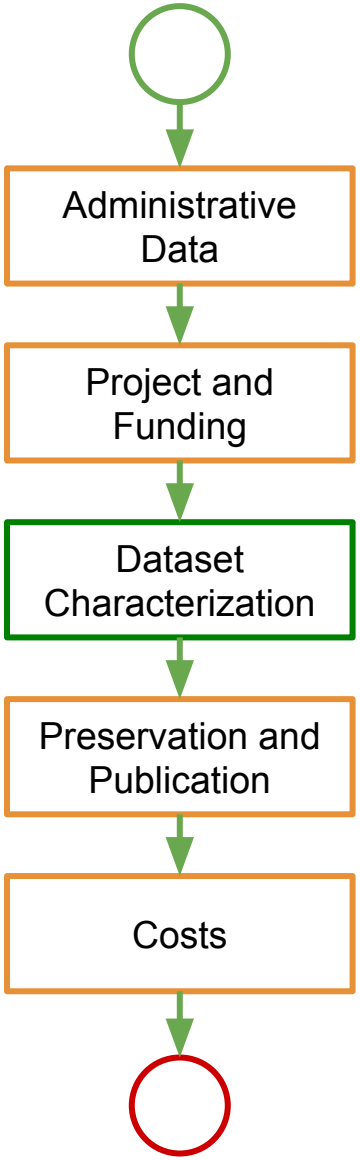
Creating a DMP (Step 3)

- **General information** – **no pitfalls here; just identify and describe the datasets**

- **Security and privacy**:
  - Which datasets include sensitive data (if any)?
  - Can they be made safe for publication (if so, how?) or should they remain private?
  - If private, then what are the access policies and how are they enforced (security)?

- **Technical resources** – include both hardware and software that were involved in data acquisition/processing.

- **Metadata**:
  - Are there established metadata practices/standards for the types of data in the datasets?
  - Are there recommended ontologies?



Creating a DMP (Step 3)

In the project

General information about all **datasets**.

Expected Data & Metadata Outputs:

1. Sample Collection:
   - Patient clinical data (< 1 MB)
   - Sample identification table (< 1 MB)
2. Phage X sequencing
   - Raw FASTQ sequencing data - NextSeq (60 MB)
   - Sample preparation & sequencing metadata - NextSeq (< 1 MB)
   - Raw FASTQ sequencing data - MinION (1 GB)
   - Sample preparation & sequencing metadata - MinION (< 1 MB)
   - Assembled Phage X genome (< 1 MB)
   - Assembly metadata (< 1 MB)
3. 16S sequencing
   - Raw FASTQ sequencing data (15 GB)
   - Sample preparation & sequencing metadata - NextSeq (< 1 MB)
   - Biome tables (< 1 MB)
4. Metatranscriptomics
   - Raw FASTQ sequencing data (120 GB)
   - Sample preparation & sequencing metadata - NextSeq (< 1 MB)
   - RNAseq count tables (< 1 MB)
   - Differential expression test results (< 1 MB)

Creating a DMP (Step 3)

In the project

Sample Collection:

In the sample collection activity, we will define a study group of volunteer disease X patients, numbering no less than 20, and a control group comprising their close relatives, 1-2 per patient. We will collect stool samples from each of the volunteers.

Any **security and privacy** issues regarding these datasets?

Creating a DMP (Step 3)

In the project

Phage X sequencing:

In the Phage X sequencing activity, we will carry out DNA sequencing of the stool samples and assemble the genome of Phage X. In order to facilitate the assembly while enabling the reliable identification of sequence variants, we will combine the higher quality but short read sequencing technology of the Illumina NextSeq 500 sequencer with the long read but lower quality technology of the Nanopore MinION sequencer.

Technical resources associated with the raw data.

But we're missing the data analysis software!!!

Creating a DMP (Step 3)

In the project

Expected Data & Metadata Outputs:

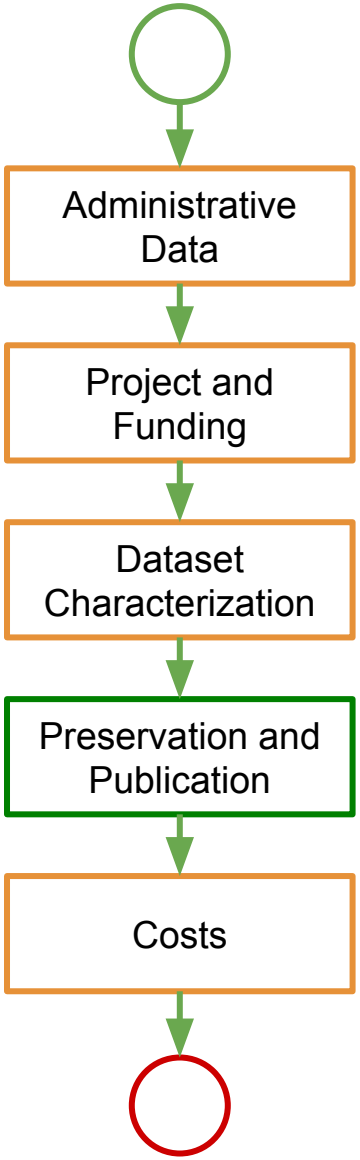
1. Sample Collection:
   - Patient clinical data (< 1 MB)
   - Sample identification table (< 1 MB)
2. Phage X sequencing
   - Raw FASTQ sequencing data - NextSeq (60 MB)
   - Sample preparation & sequencing metadata - NextSeq (< 1 MB)
   - Raw FASTQ sequencing data - MinION (1 GB)
   - Sample preparation & sequencing metadata - MinION (< 1 MB)
   - Assembled Phage X genome (< 1 MB)
   - Assembly metadata (< 1 MB)
3. 16S sequencing
   - Raw FASTQ sequencing data (15 GB)
   - Sample preparation & sequencing metadata - NextSeq (< 1 MB)
   - Biome tables (< 1 MB)
4. Metatranscriptomics
   - Raw FASTQ sequencing data (120 GB)
   - Sample preparation & sequencing metadata - NextSeq (< 1 MB)
   - RNAseq count tables (< 1 MB)
   - Differential expression test results (< 1 MB)

Some (not exhaustive) **metadata**.

We're missing the standards which these metadata will follow!!!

Creating a DMP (Step 4)

- **Step 4 - Preservation and Publication**
  - **Characterization** of the **preservation** and **publication** policies **for each** of the identified **datasets.**
  - The information is classified in three sections:
    - Information regarding the **policies** on how each dataset is **distributed**.
    - Information on the data **host for each** of the identified **distributions**.
    - Characterization of the **licenses** associated with each **distribution policies**.



Creating a DMP (Step 4)

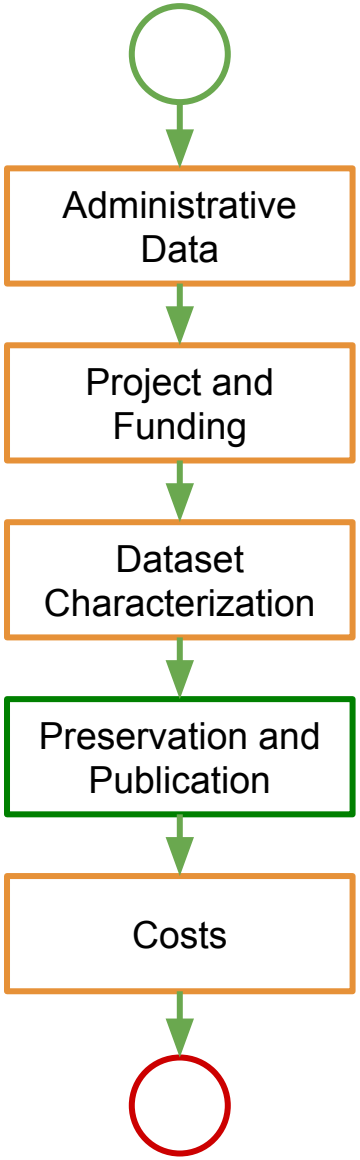
In the project

??????

This is not usually covered in project descriptions.

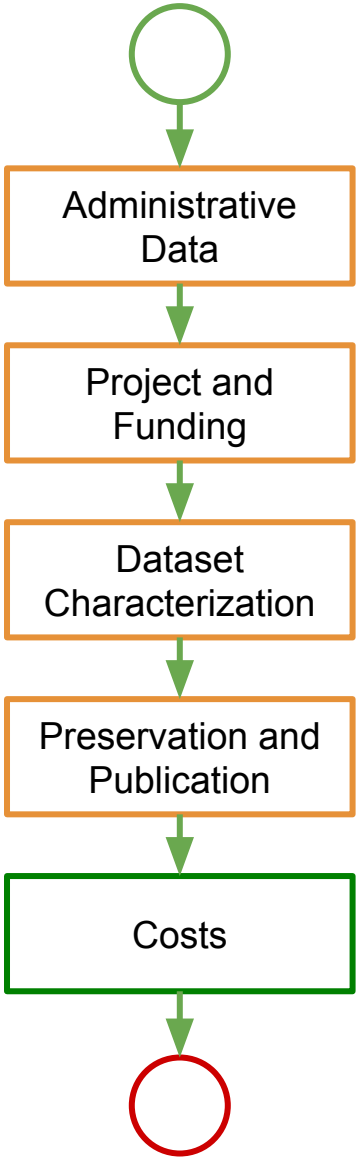
Creating a DMP (Step 4)

- **Distribution policies (for each dataset):**
  - Is the dataset going to be published in a public repository, a repository with restricted access, or will it remain fully private?
  - Is it going to feature in a scientific publication?
- **Host for each distribution**:
  - The public repository in question OR the institute hosting the repository or server where the dataset is hosted.
- **License** for each distribution:
  - Do you want to be credited by users of your data (attribution)?
  - Do you want to allow the data to be used commercially?



Creating a DMP (Step 5)

- **Step 5 - Costs**
  - **Characterization** of the **costs** associated with this DMP.
    - The numeric value associated with each cost (a rough estimate is fine).
- **Costs should include:**
  - **Staff** directly involved in any stage of the data lifecycle (e.g. acquisition, analysis, management, publication, storage).
  - **Hardware and software** required at any stage of the data lifecycle



Creating a DMP (Step 5)

In the project

Expected Data & Metadata Outputs:

1. Sample Collection:
   - Patient clinical data (< 1 MB)
   - Sample identification table (< 1 MB)
2. Phage X sequencing:
   - Raw FASTQ sequencing data - NextSeq (60 MB)
   - Sample preparation & sequencing metadata - NextSeq (< 1 MB)
   - Raw FASTQ sequencing data - MinION (1 GB)
   - Sample preparation & sequencing metadata - MinION (< 1 MB)
   - Assembled Phage X genome (< 1 MB)
   - Assembly metadata (< 1 MB)

What other costs would you consider?

We can infer **storage costs** here.

But there are other costs to consider!!!

Thank you!

BioData.pt

E-mail: info@biodata.pt
www.biodata.pt

Co-financiado por: