

Minority Report: A Execução do Plano de Gestão de Dados em Análise



Daniel Agostinho^a, João Pereira^a, Alexandre Sayal^a,

Bruno Direito^b

bruno.direito@uc.pt

a Centro de Imagem Biomédica e Investigação Translacional (CIBIT),
Universidade de Coimbra

b Centro de Informática e Sistemas da Universidade de Coimbra



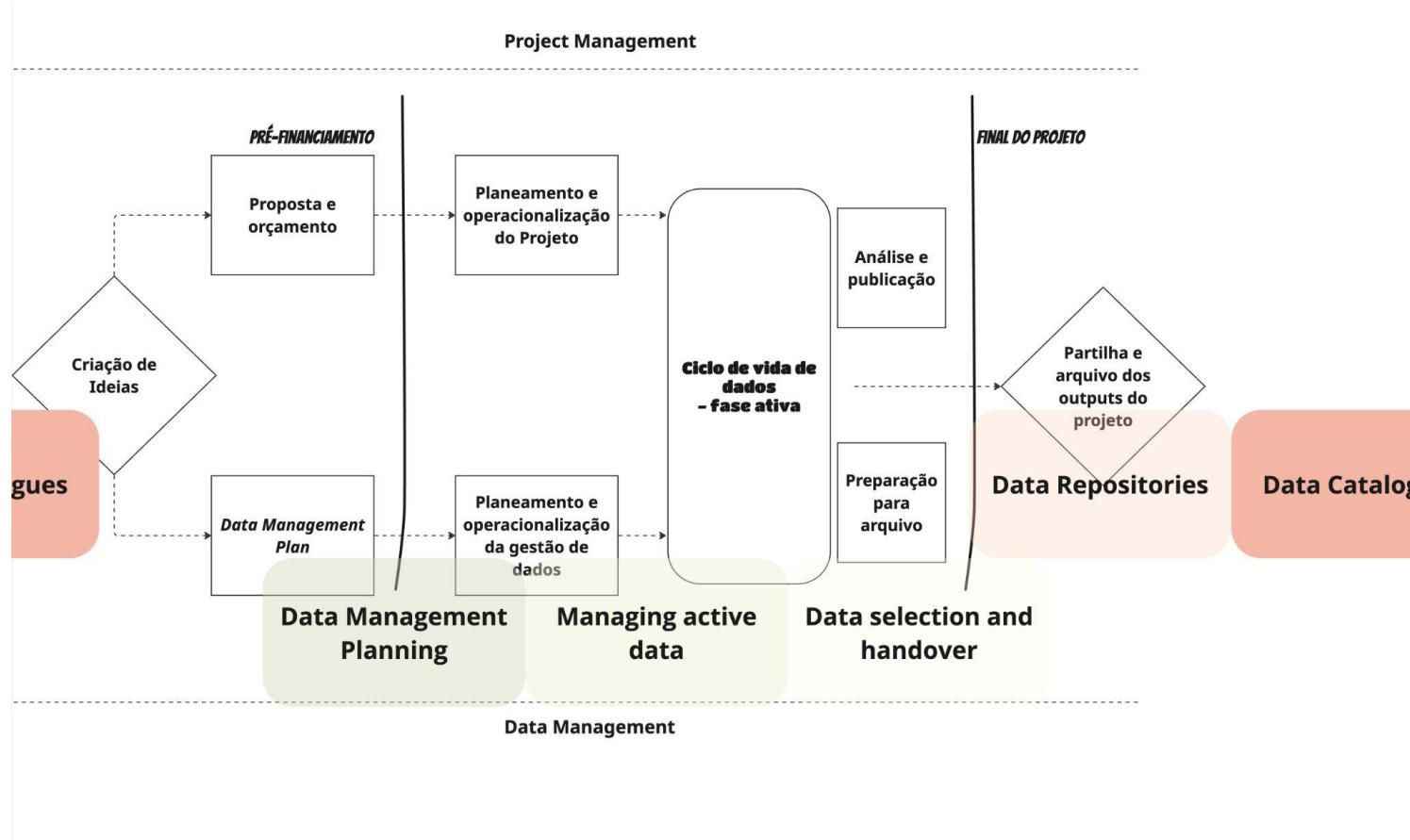
This work is licensed under CC BY SA. To view a copy of this license, visit
<https://creativecommons.org/licenses/by-sa/4.0/>



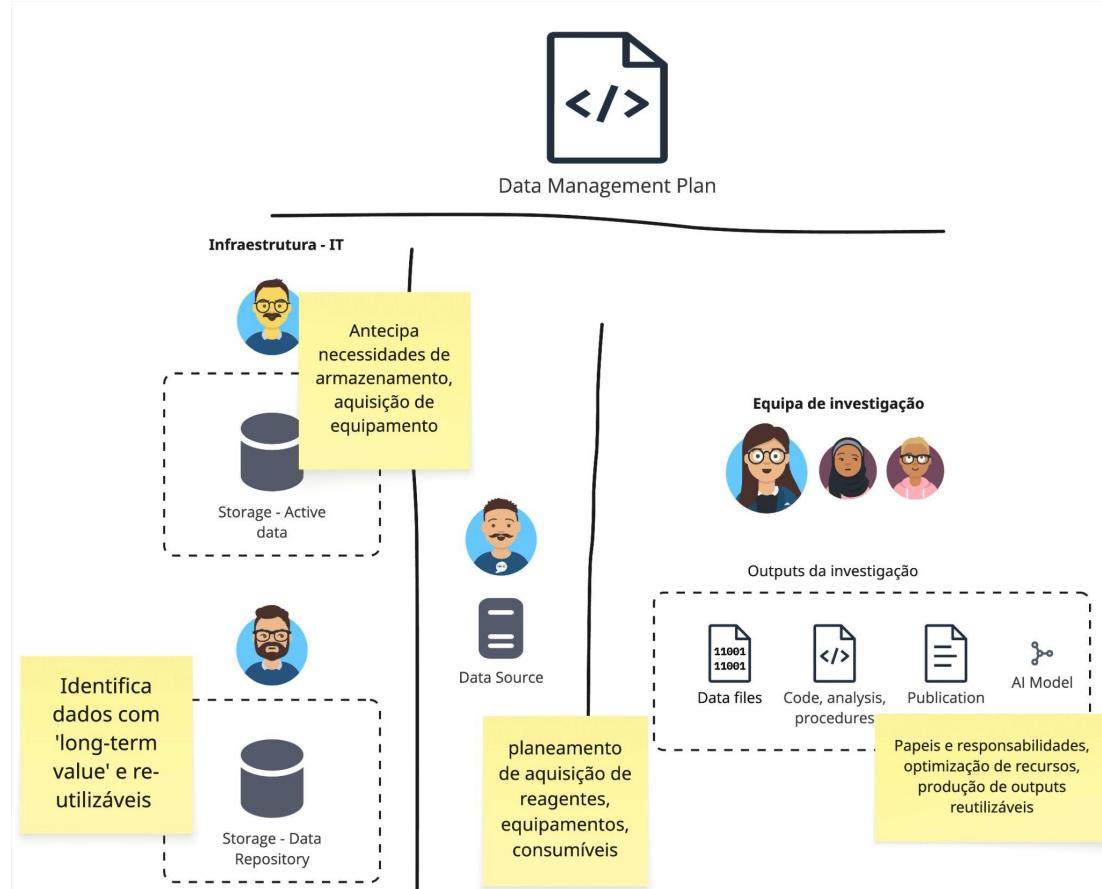


- *Minority report*

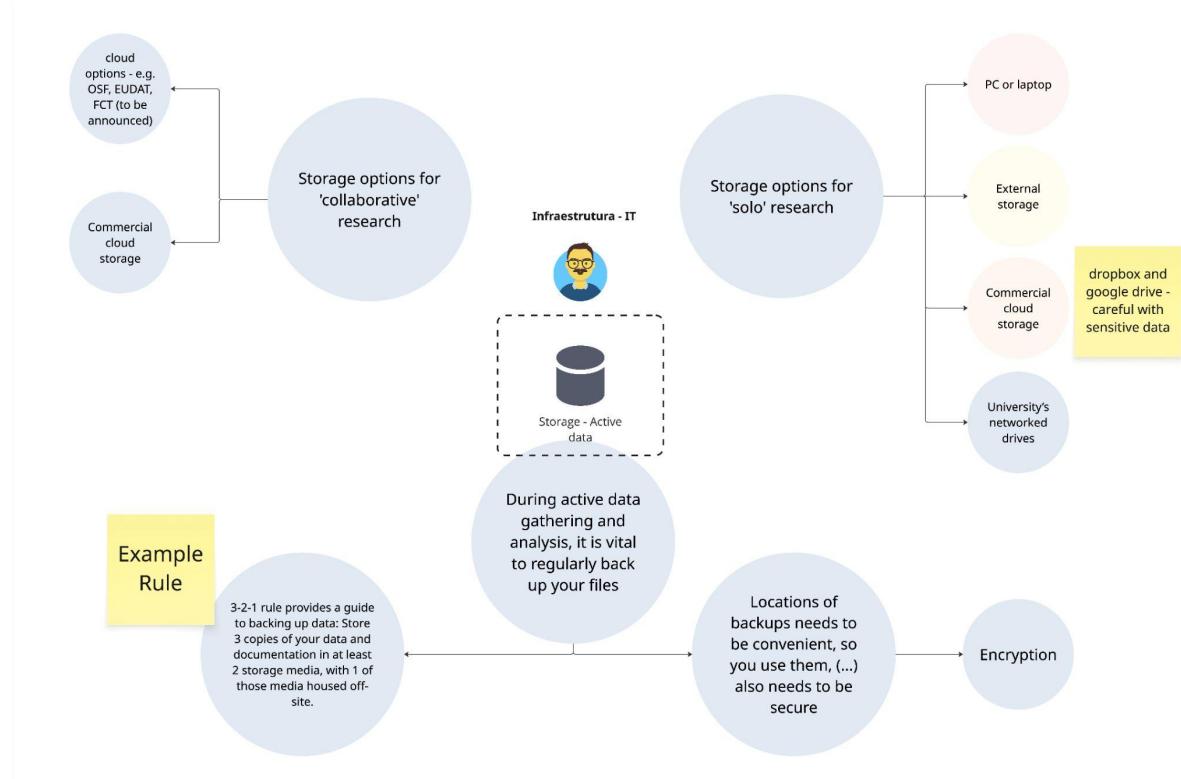
- Em Washington, no ano de 2054, o crime foi erradicado graças à divisão 'Pré-Crime'. O sistema baseia-se em três 'Precogs' que visualizam assassinatos (**fragmentos desestruturados de informação**) antes de estes ocorrerem.
- O chefe da unidade, John Anderton, **interpreta e processa estes fragmentos de dados** para localizar e prender os culpados.
- Descobre a existência de 'Relatórios Minoritários', que são sistematicamente apagados para **manter a ilusão de que o sistema é infalível**.



A importância do planeamento



Requisitos Core PGD

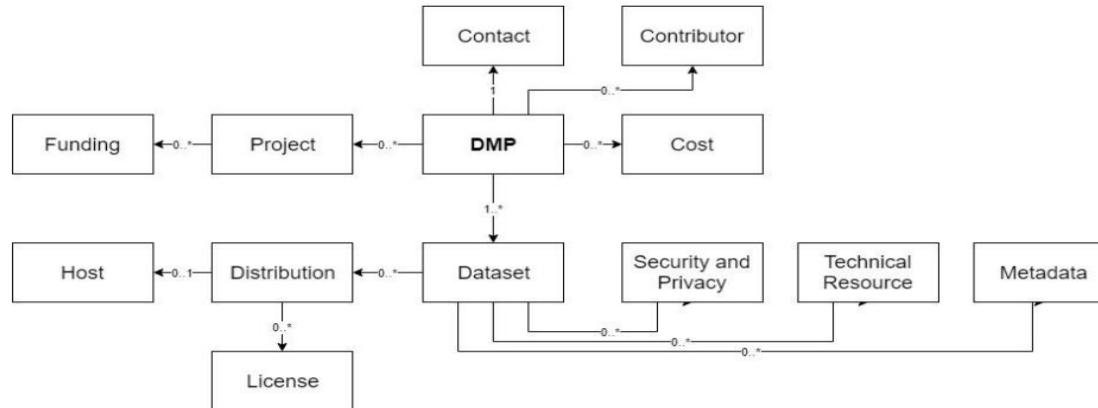


Plano de Gestão de Dados como um instrumento do **projeto**

- Para ser **prático e útil** para o investigador e instituições de acolhimento, o PGD deve ser:
 - Um **documento vivo**, atualizado sempre que necessário
 - Acionável por máquina
 - Comply com standards claros e aceites pela comunidade
 - Ser partilhado
 - ter uma versão e um DOI associado

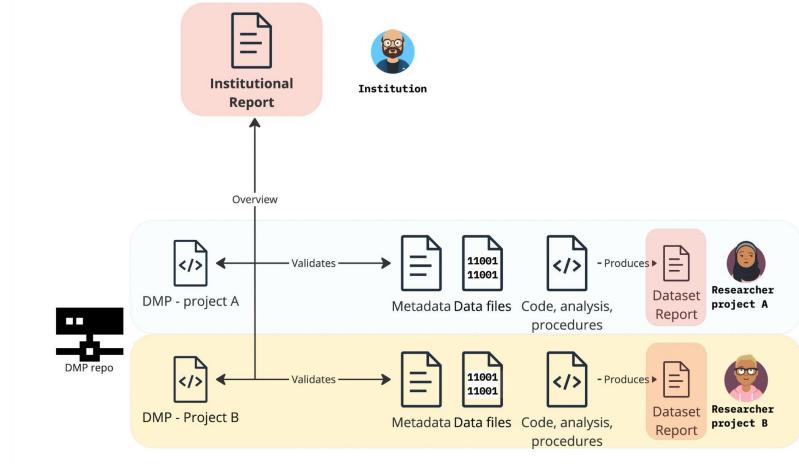
Plano de Gestão de Dados como um instrumento do projeto e de uma instituição

- Um conjunto mínimo de elementos no documento, com termos universais e transversais a todos os templates para assegurar a interoperabilidade e integração

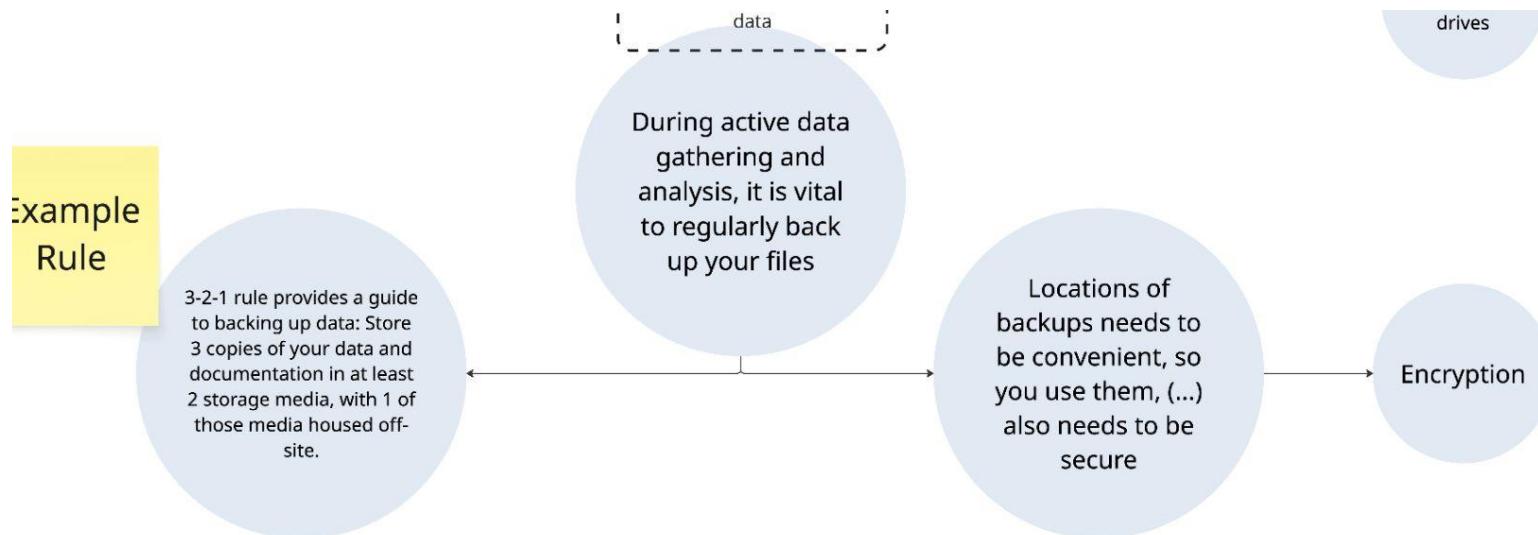


Plano de Gestão de Dados como um instrumento do projeto e de uma instituição

- Os benefícios do Plano de Gestão de Dados
 - Promove boas práticas de gestão de dados
 - Orienta a compliance com os princípios FAIR
 - Assegura a alocação de recursos para as atividades de gestão de dados.
- maDMPs:
 - **Automation (creation, validation, policy enactment)**
 - **Aumenta a utilidade (individual e institucional) do documento**

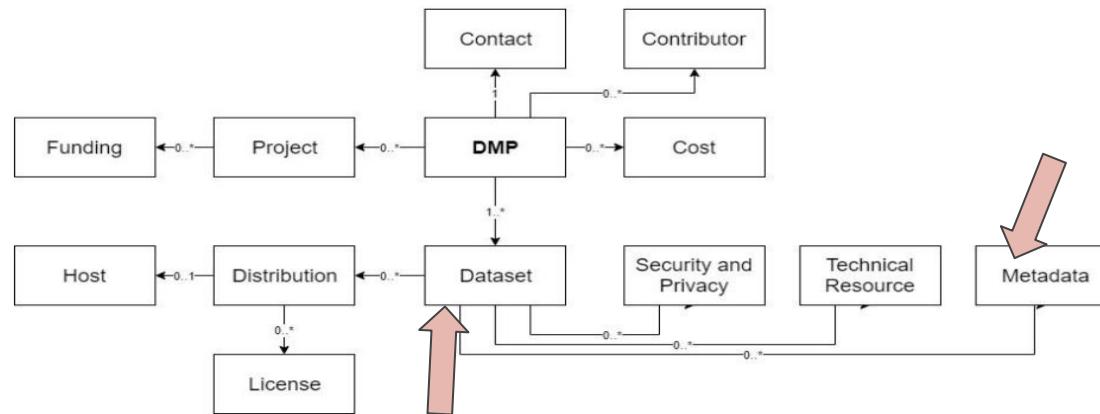


Armazenamento e backup (active data phase), PGD como ferramenta de monitorização



Plano de Gestão de Dados como um instrumento do projeto e de uma instituição

- Em grande parte das áreas científicas, o PGD detém informação relativamente
 - ao número de elementos (dados)
 - o seu volume
 - a sua estrutura (standard)



Plano de Gestão de Dados como um instrumento do projeto e de uma instituição

- ARGOS
 - FCT template

1.4. Organizations
1.5. Language
1.6. Contact
1.7. Authors
2. Funding
3. License
4. Templates
4.1. FCT - Template in English ✓
4.1.1. DATA INFORMATION
4.1.1.1. What existing data will be re...

4.1.1.3. What is the data volume?

4.1.1.3.1. How much data storage will your project require in total? ✓

Volume estimate*
0 - 10 GB

If relevant please justify the volume estimate
40 participants; 9 photos per participant; 4MB per photo;

1.4. Organizations
1.5. Language
1.6. Contact
1.7. Authors
2. Funding
3. License
4. Templates
4.1. FCT - Template in English ✓
4.1.1. DATA INFORMATION
4.1.2. DOCUMENTATION AND METADATA
4.1.2.1. Documentation
4.1.2.1.1. Indicate what documentation ...
4.1.2.2. Metadata
4.1.3. STORAGE AND SECURITY OF DA...

4.1.2.2. Metadata

4.1.2.2.1. Is there metadata associated with the data? ✓

Yes No Required

4.1.2.2.2. Indicate which metadata will be provided to help others identify and discover the data. ✓

To be findable, accessible, interoperable and reusable, data must be accompanied with descriptive information in the form of metadata.

Please specify**
Date, time, GPS coordinates (if enabled by user), camera model, and settings.

Plano de Gestão de Dados como um instrumento do projeto e de uma instituição



Identificação dos elementos chave - numero de elementos, standard e tamanho

```
# The specific fieldId you are searching for
target_field_id = '18df1620-7138-51ca-3b6a-e2db5f34837'

[4] ✓ 0.0s

# XPath to find the <value> sibling of the <item> that contains the correct
<fieldId>
# Breaking it down:
# //item...
# //fieldId/text() = '...'           - Find an <item> element anywhere...
# //item/...                          - ...where its child <fieldId> has the specified
text
# //following-sibling::value          - Then, get that <item>'s sibling named <value>.
xpath_query = "//item[fieldId/text() = '{target_field_id}']"

# Execute the search
value_elements = root.xpath(xpath_query)

[5] ✓ 0.0s

> v
# count number of elements found
print("Number of elements found: {}".format(len(value_elements)))

# get the parent of the found elements
for elem in value_elements:
    # get children of elem and print their tags and text
    for child in elem:
        # print("Tag: {} Text: {}".format(child.tag, child.text))

        if child.tag == 'textValue':
            print("Text Value: {}".format(child.text))

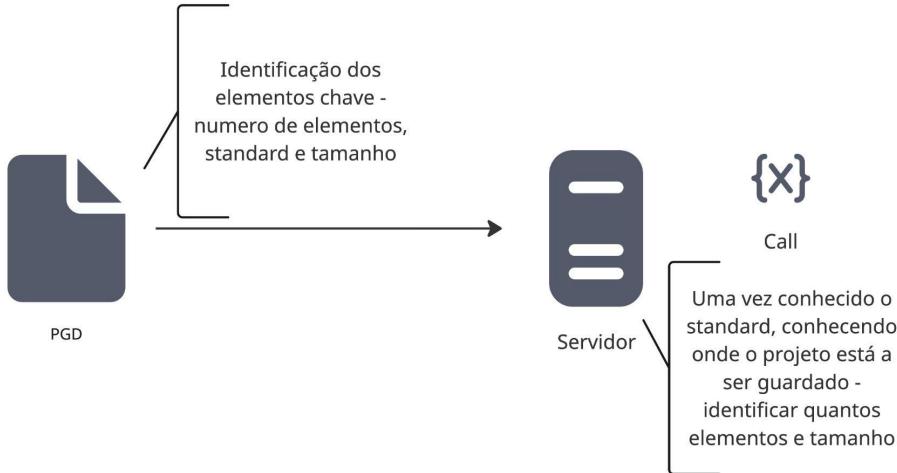
            # parse text and identify the text representing the number of
            # participants
            text = child.text
            if text:
                # split by spaces and look for the first integer
                parts = text.split()
                for part in parts:
                    if part.isdigit():
                        print("Number of participants: {}".format(part))
                        break

[6] ✓ 0.0s

...
Number of elements found: 1
Text Value: We aim to acquire data from approximately 50 participants - each participant requires 2 GB
Number of participants: 50
```

PGD

Plano de Gestão de Dados como um instrumento do projeto e de uma instituição



```
# Here we're using an example BIDS dataset that's bundled with the pybids tests
data_path = os.path.join('/Volumes/T7/datasets/BIDS-BRAINPLAYBACK-TASK1')

# Initialize the layout
layout = BIDSLayout(data_path)

# Print some basic information about the layout
layout
[5] ✓ 0.7s
BIDS Layout: ...assets/BIDS-BRAINPLAYBACK-TASK1 | Subjects: 12 | Sessions: 12 | Runs: 48

# get the number of sessions per participant
sessions = layout.get_sessions()
n_sessions = len(sessions)

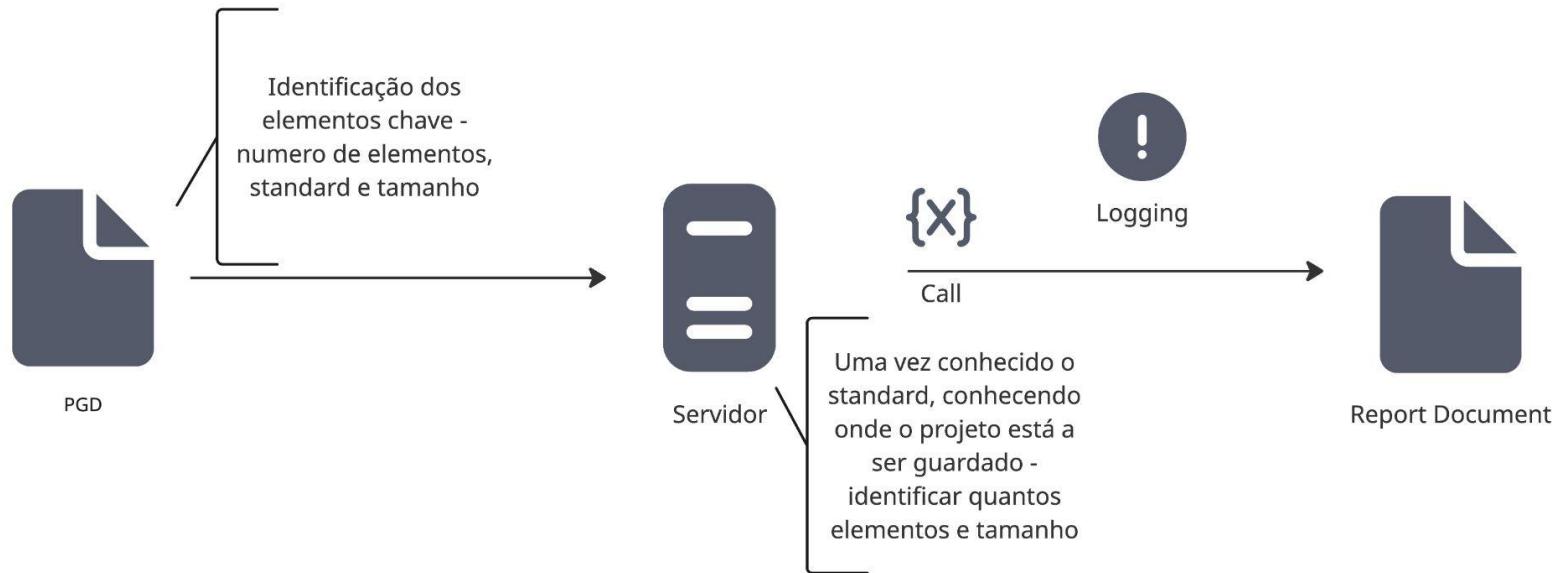
# for each participant, get the number of runs
runs = layout.get_runs()
n_runs = len(runs)

# get the number of participants
participants = layout.get_subjects()
n_participants = len(participants)

# print the results
print(f'Number of participants: {n_participants}')
print(f'Number of sessions: {n_sessions}')
print(f'Number of runs: {n_runs}')

[6] ✓ 0.1s
Number of participants: 12
Number of sessions: 1
Number of runs: 4
```

Plano de Gestão de Dados como um instrumento do projeto e de uma instituição



```

projected_participants = 50
projected_size = 100
projected_size_ongoing = total_size * (projected_participants / n_participants)
print('Projected size for (projected_participants) participants:
      (projected_size_ongoing / (1024 ** 3))f GB')

# create a plot with the % of participants already available

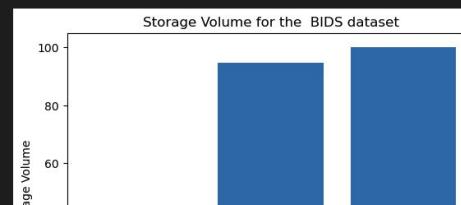
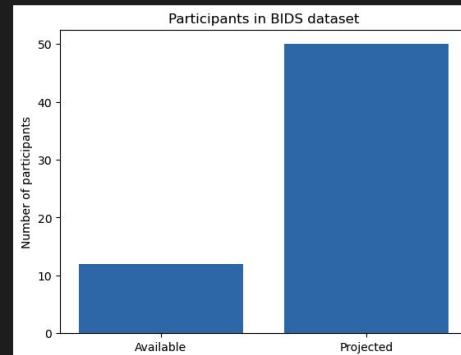
plt.bar(['Available', 'Projected'], [n_participants, projected_participants])
plt.ylabel('Number of participants')
plt.title('Participants in BIDS dataset')
plt.show()

# create plot with % of size occupied in GB
plt.bar(['Occupied', 'Projected', 'Asked'], [total_size / (1024 ** 3),
      projected_size_ongoing / (1024 ** 3), projected_size])
plt.ylabel('Storage Volume')
plt.title('Storage Volume for the BIDS dataset')
plt.show()

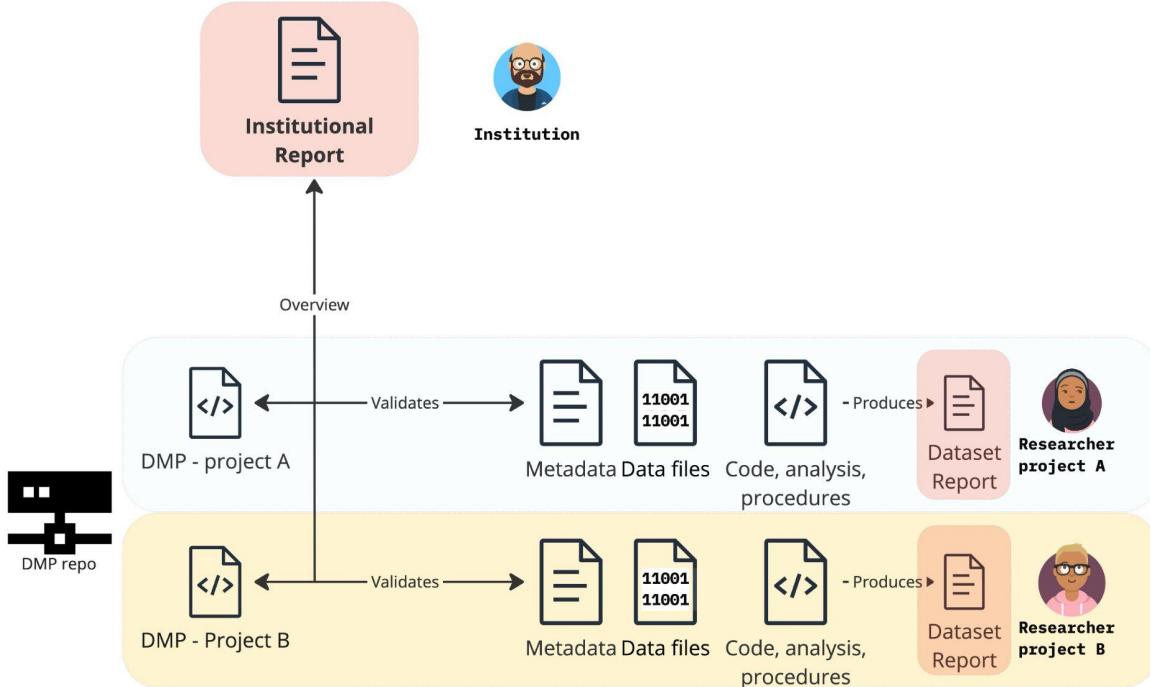
```

✓ 0.0s

Projected size for 50 participants: 94.62 GB

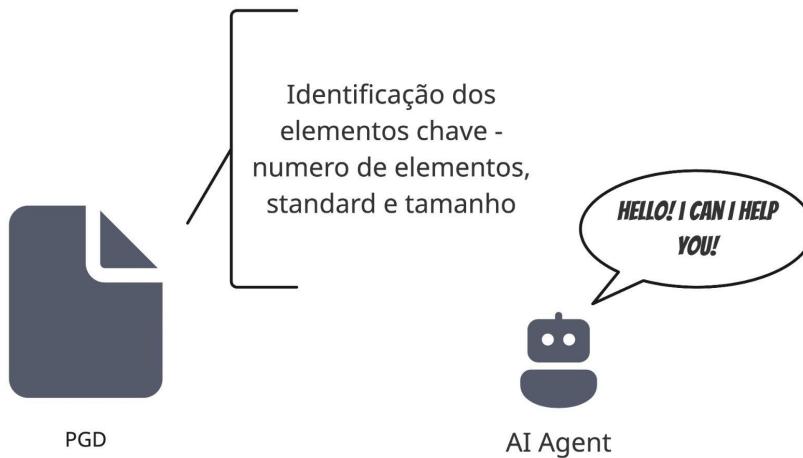


- Relatório com **taxa de execução do projeto** - *data management*
 - número de elementos
 - tamanho

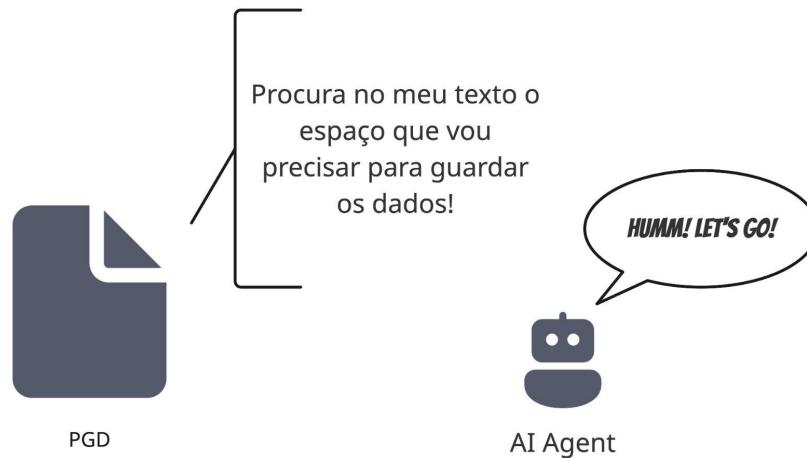


- **Visão institucional** do estado atual dos projetos, recursos/serviços de armazenamento
 - optimização, antecipação de custos, etc.

Plano de Gestão de Dados como um instrumento do **projeto e de uma instituição** **- Agentes de IA baseados em LLM**



Plano de Gestão de Dados como um instrumento do **projeto e de uma instituição** **- Agentes de IA baseados em LLM**



Plano de Gestão de Dados como um instrumento do projeto e de uma instituição

- Agentes de IA baseados em LLM

```
⬇ dmp_report.md > abc # Data Volume Report
1  # Data Volume Report
2
3  ## Total Estimated Size
4  **10 GB**
5
6  ### Breakdown
7  * Estimated storage requirement: 0 - 10 GB
8  * Calculated from photo generation: 1.44 GB (40 participants * 9 photos/participant * 4MB/photo)
9
10 ### Evidence
11 > "1.3 What is the data volume?
12 > 1.3.1 How much data storage will your project require in total?
13 > 0 - 10 GB
14 > 40 participants; 9 photos per participant; 4MB per photo;"
15
16 ### Notes
17 * The total estimated size is reported as the upper bound of the stated range (0 - 10 GB) for total storage requirements.
18 * The explicit calculation for the photo generation results in approximately 1.44 GB.
```

Q&A!

or later at bruno.direito@uc.pt