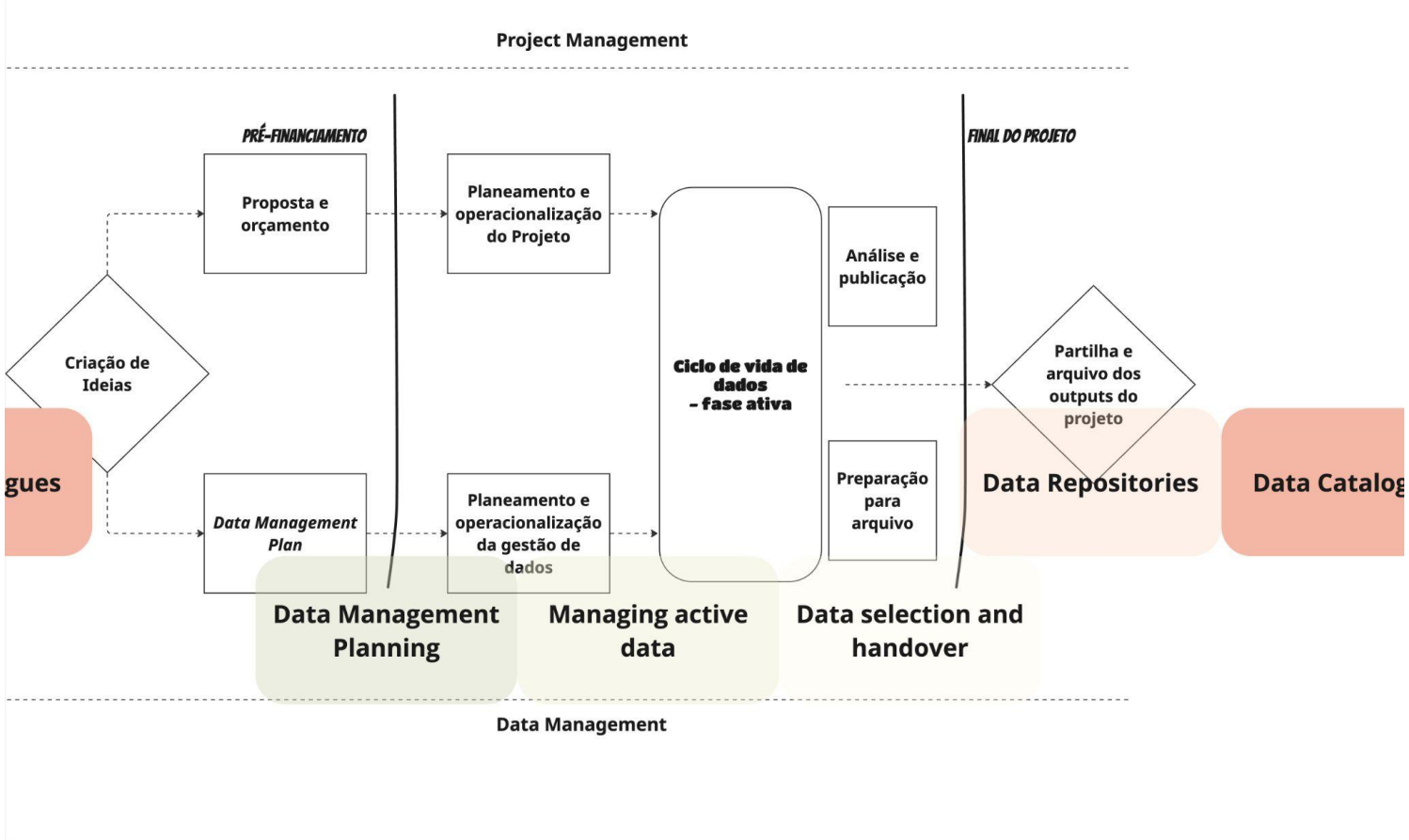
Minority Report: A Execução do Plano de Gestão de Dados em Análise

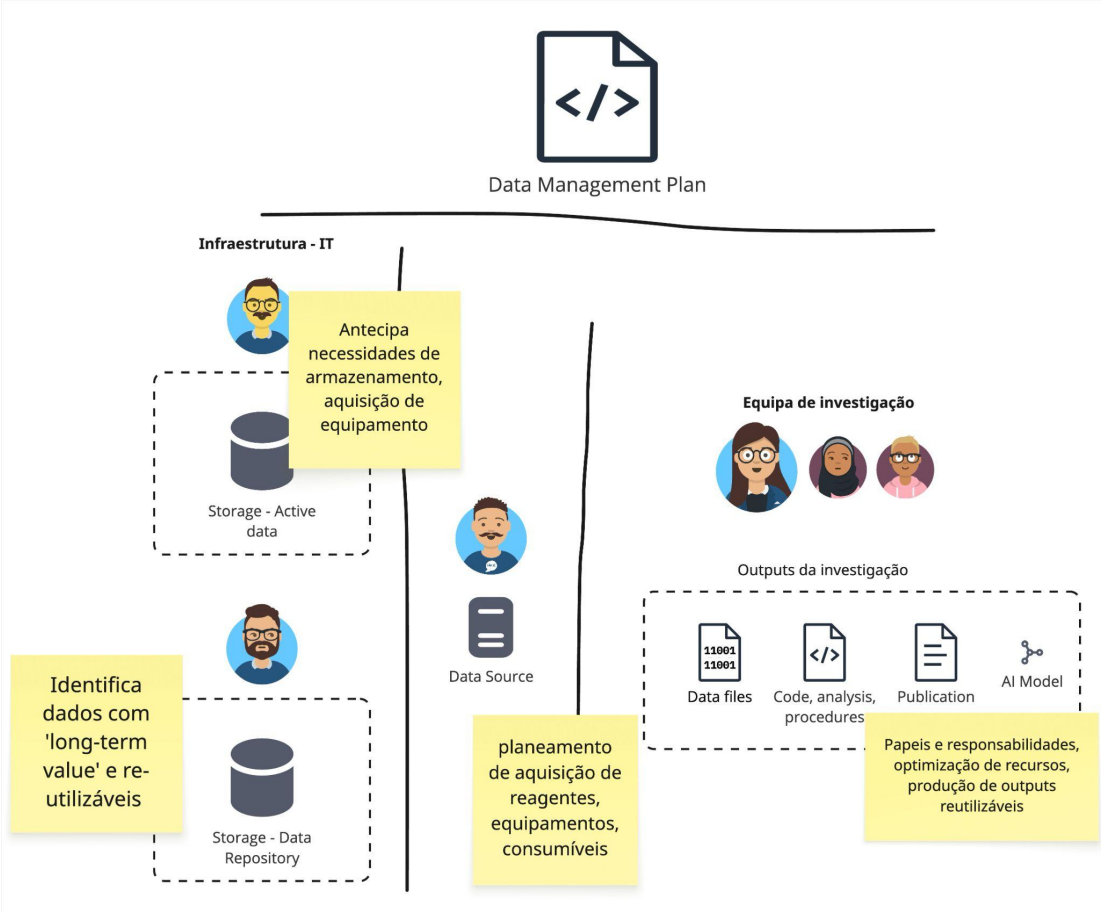
Daniel Agostinho[a], João Pereira[a], Alexandre Sayal[a],
Bruno Direito[b]
bruno.direito@uc.pt

a Centro de Imagem Biomédica e Investigação Translacional (CIBIT), Universidade de Coimbra
b Centro de informática e Sistemas da Universidade de Coimbra

- *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**.

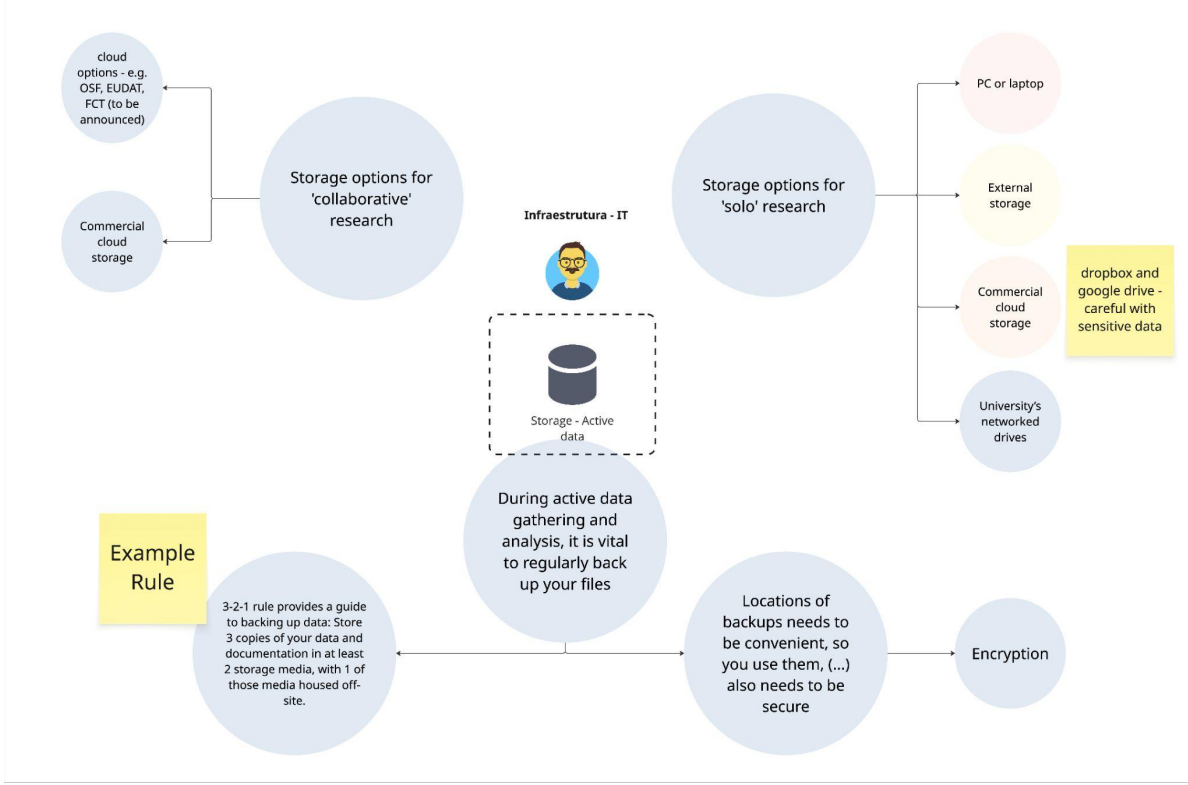


A importância do planeamento



Requisitos *Core PGD*

Armazenamento e backup durante o projeto

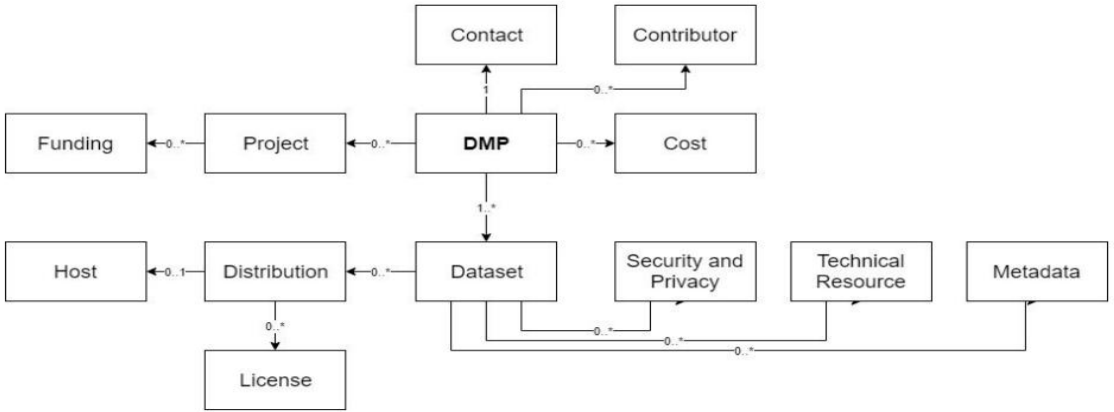


Plano de Gestão de Dados como um instrumento do **projeto**

- 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

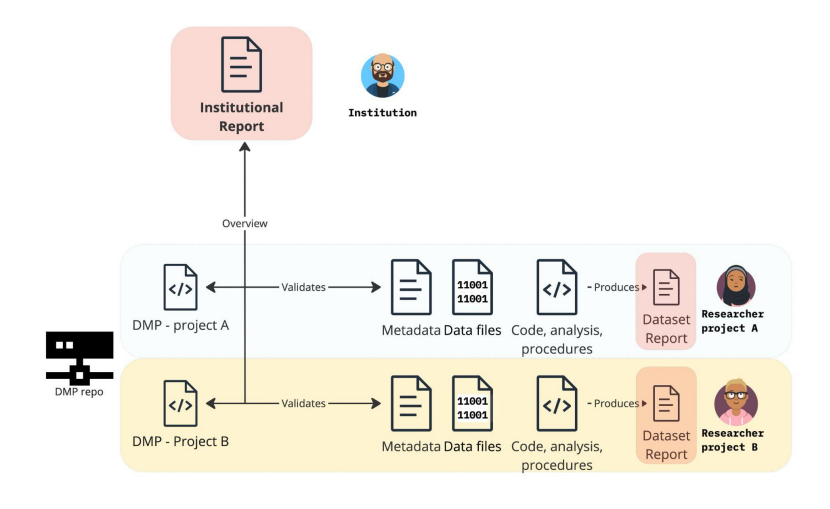
Plano de Gestão de Dados como um instrumento do **projeto e de uma instituição**

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

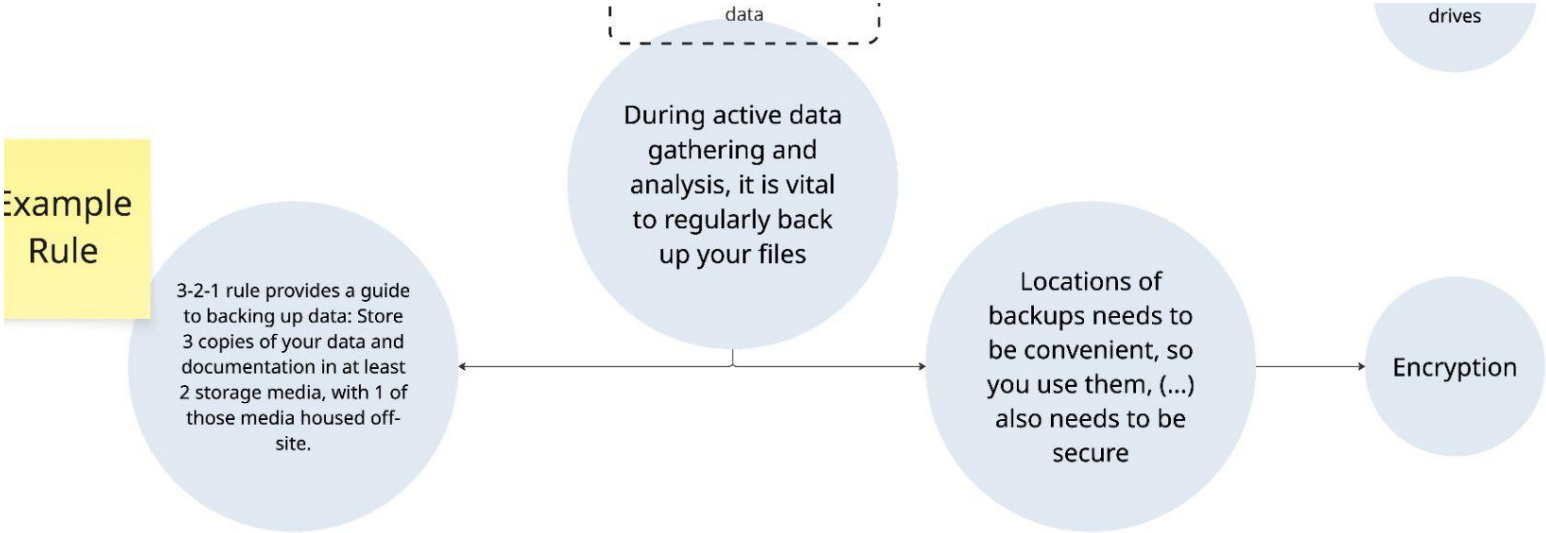


Plano de Gestão de Dados como um instrumento do **projeto e de uma instituição**

- 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**

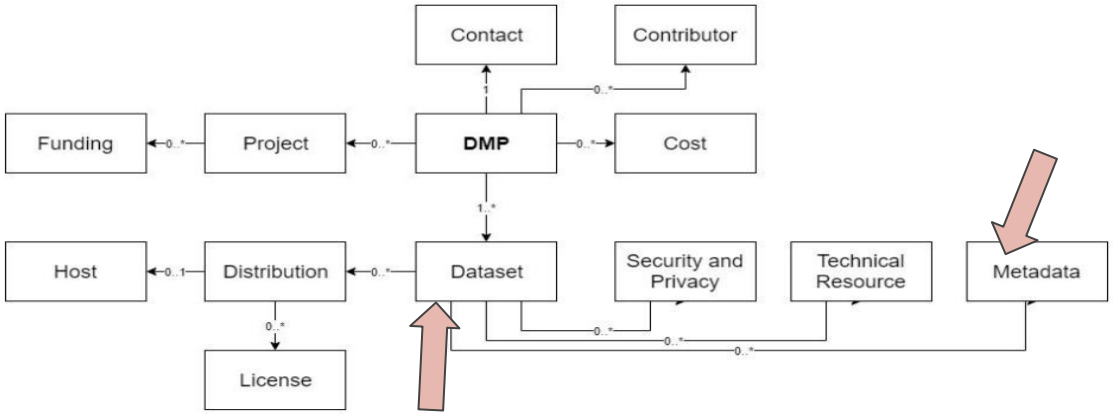


Armazenamento e backup (*active data phase*), PGD como ferramenta de monitorização



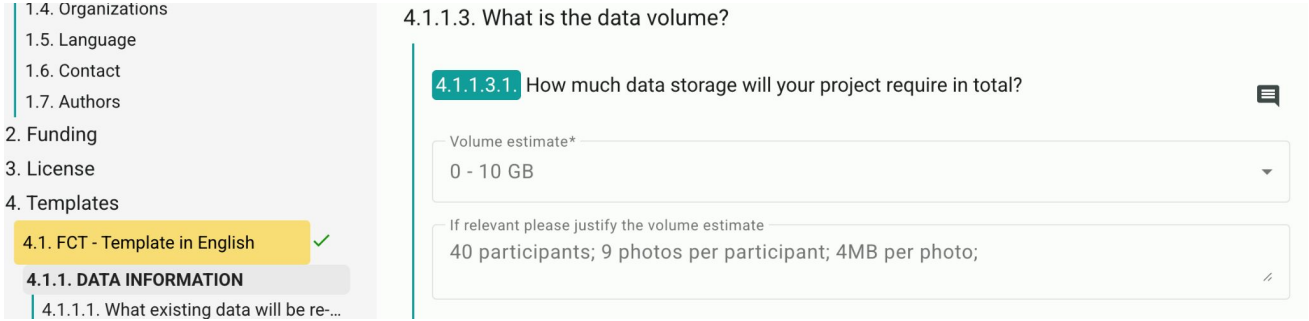
Plano de Gestão de Dados como um instrumento do **projeto e de uma instituição**

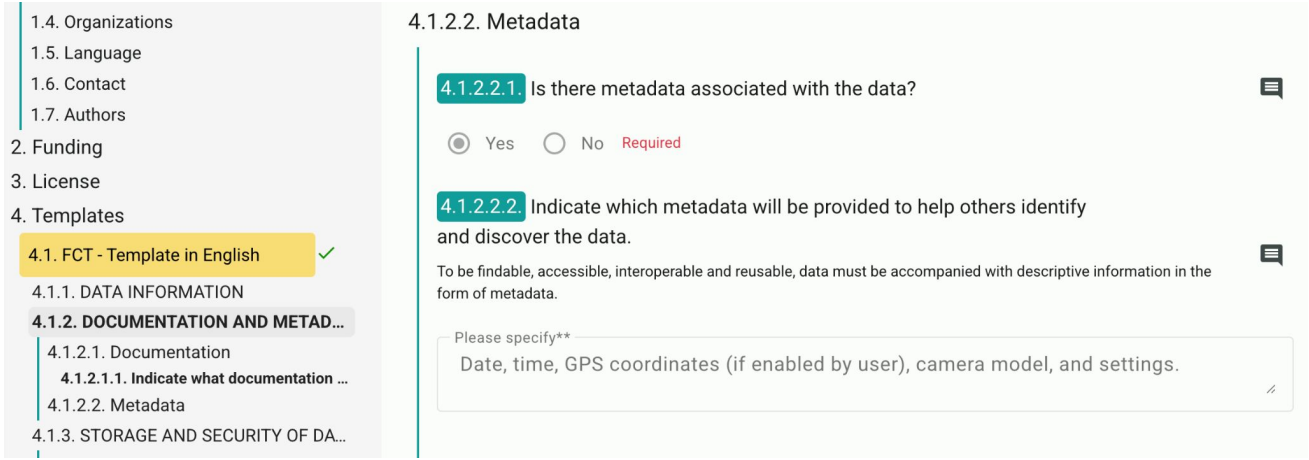
- Em grande parte das áreas científicas, o PGD detém informação relativamente
  - ao número de elementos (dados)
  - o seu volume
  - a sua estrutura (standard)



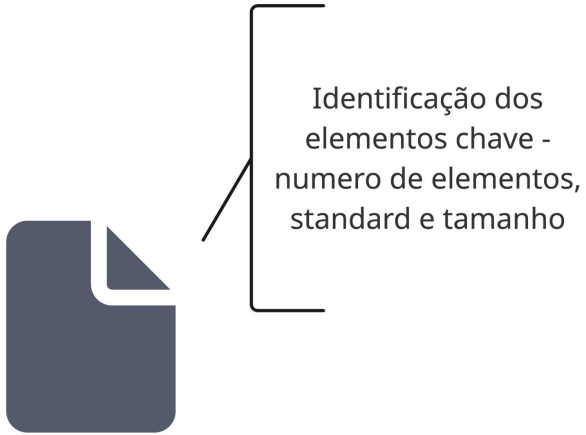
Plano de Gestão de Dados como um instrumento do **projeto e de uma instituição**

- ARGOS
  - FCT template





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



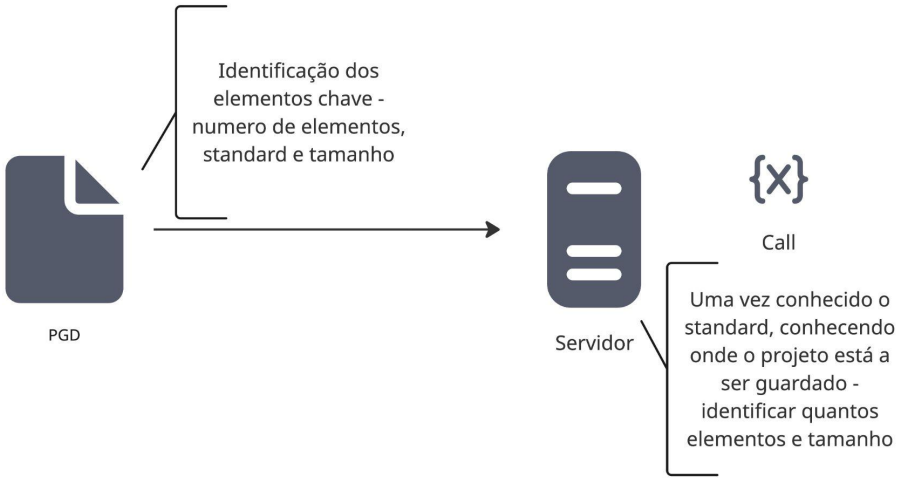

```
# The specific fieldId you are searching for
target_field_id = '18df1620-7138-51ca-3b6a-ee2db5f34837'
```
[4] ✓ 0.0s

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

# Execute the search
value_elements = root.xpath(xpath_query)
```
[5] ✓ 0.0s

```
# count number of elements found
print(f"Number of elements found: {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(f"Tag: {child.tag}, Text: {child.text}")

        if child.tag == 'textValue':
            print(f"Text Value: {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(f"Number of participants: {part}")
                        break
```
[14] ✓ 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
```

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
```

✓ 0.7s

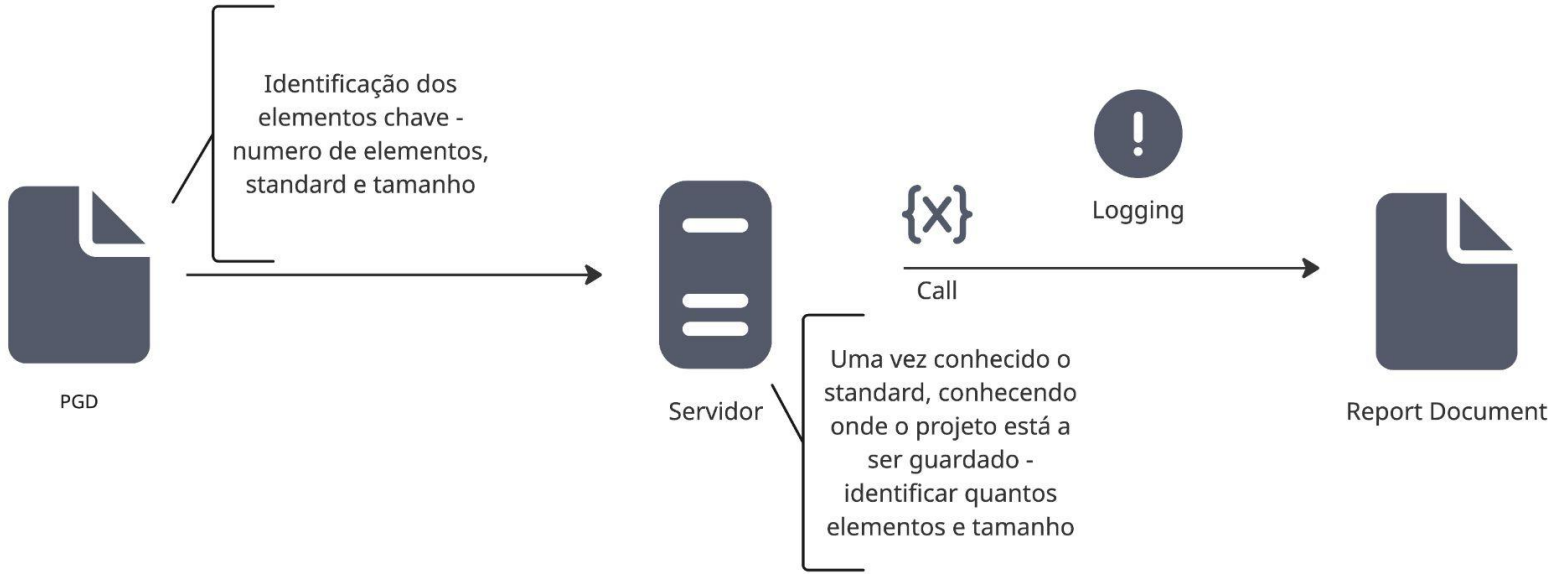
```
BIDS Layout: ...asets/BIDS-BRAINPLAYBACK-TASK1 | Subjects: 12 | Sessions: 12 | Runs: 48
```

```
# 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}')
```

✓ 0.1s

```
Number of participants: 12
Number of sessions: 1
Number of runs: 4
```

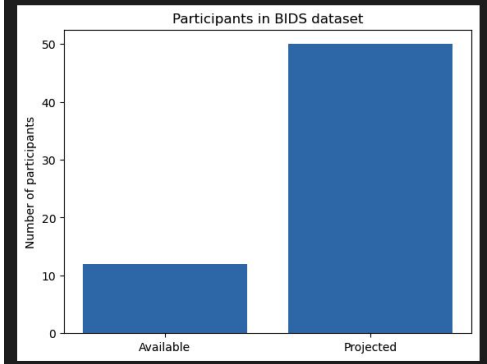
Plano de Gestão de Dados como um instrumento do **projeto e de uma instituição**

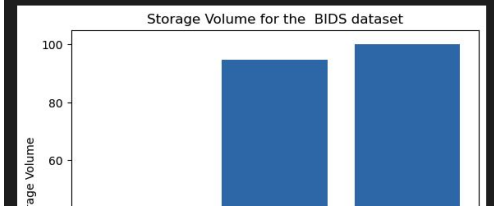


```
projected_participants = 50
projected_size = 100
projected_size_ongoing = total_size * (projected_participants / n_participants)
print(f'Projected size for {projected_participants} participants:
{projected_size_ongoing / (1024 ** 3):.2f} GB')

# 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()
```

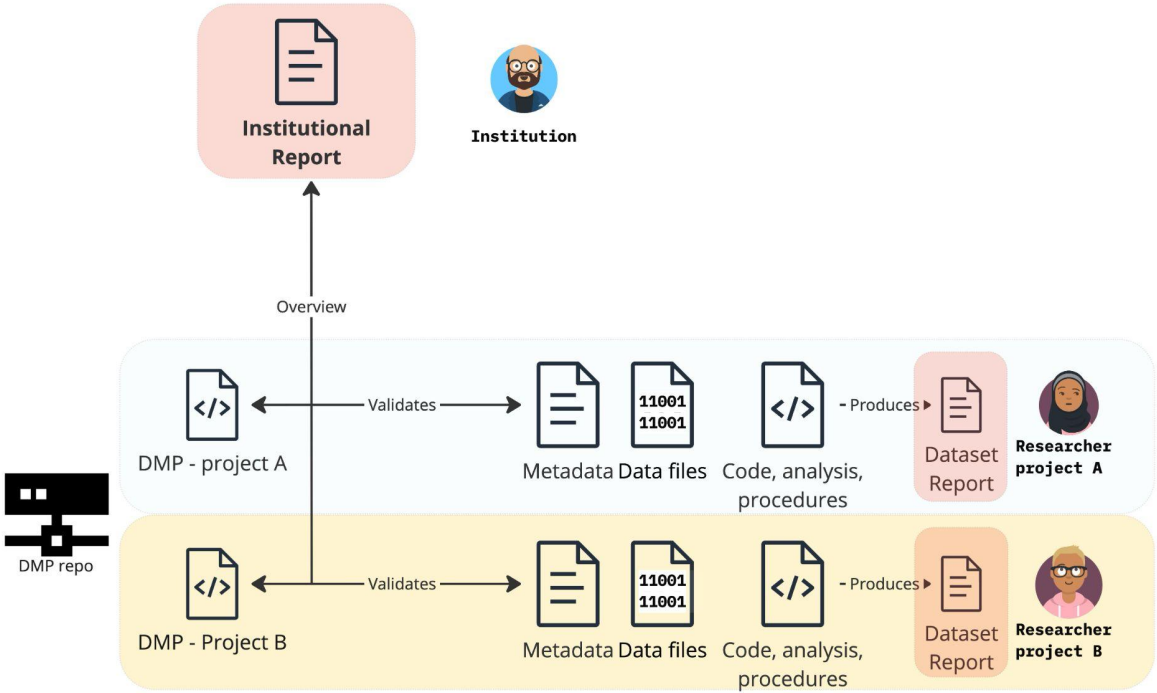
✓ 0.0s

```
Projected size for 50 participants: 94.62 GB
```





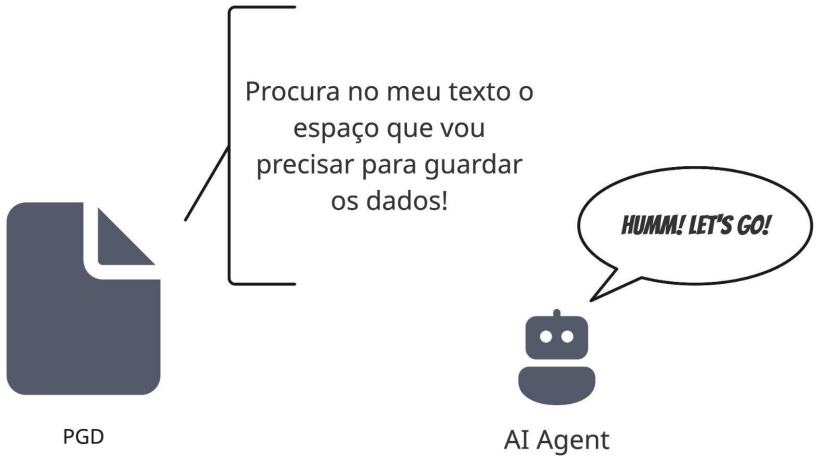
- Relatório com **taxa de execução do projeto** - *data management*
  - número de elementos
  - tamanho



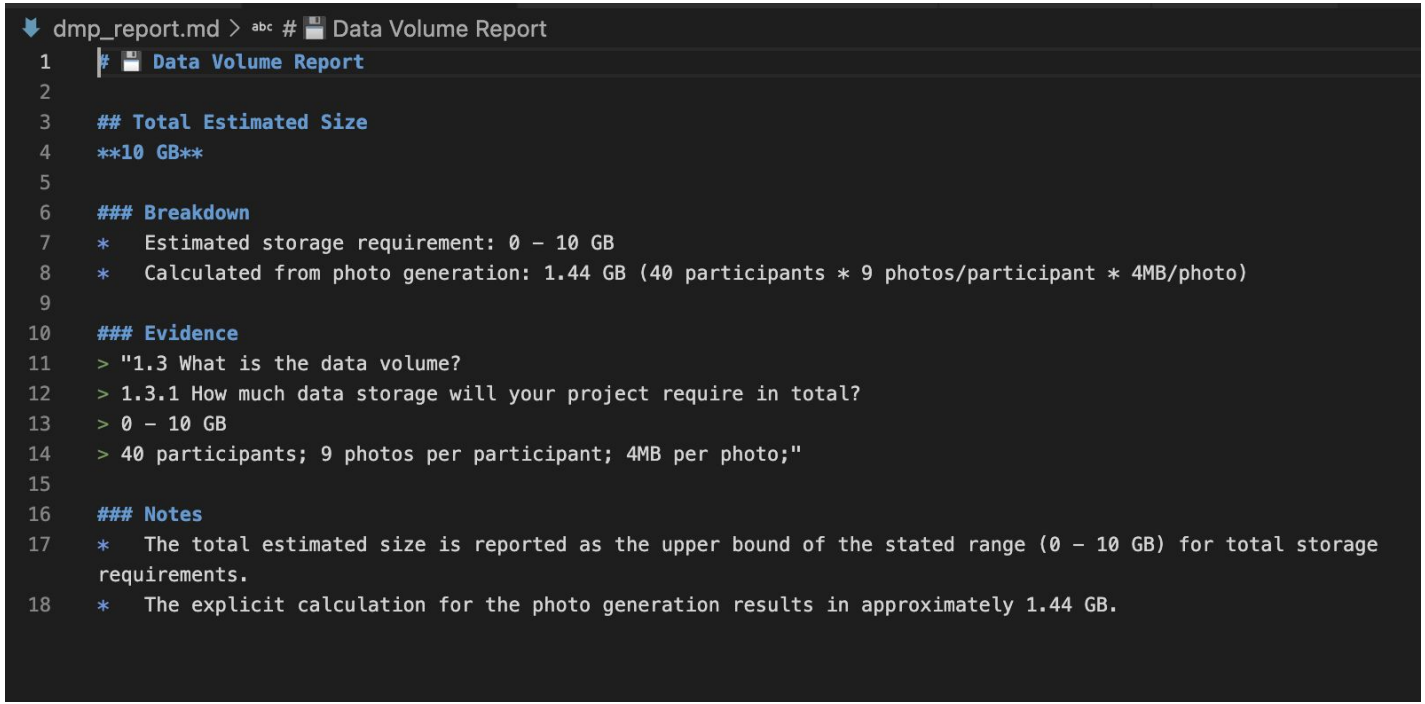
- **Visão institucional** do estado atual dos projetos, recursos/serviços de armazenamento
  - optimização, antecipação de custos, etc.

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**



Q&A!

or later at bruno.direito@uc.pt