Jornadas FCCN 2023

GESTÃO DE DADOS DE INVESTIGAÇÃO

Workshop

Boas Práticas na Gestão e Partilha de Dados de Investigação

Filipa Pereira 8A13-4EA4-A512

João Cardoso 2012-B202-C893

27 de junho 2023

Programa

- **Contexto**
  - Ciência aberta e dados abertos
  - Ciclo de vida dos dados de investigação
- **Sistema de Planos de Gestão de Dados (PGD):**
  - O que são PGD?
  - Ferramenta de elaboração de PGD - Argos
  - Demonstração prática
- **Serviço de Repositório de Dados de Investigação:**
  - Porquê depositar/preservar dados de investigação?
  - Repositório Polen - Dataverse
  - Demonstração prática

Jornadas FCCN 2023

Contexto

Filipa Pereira

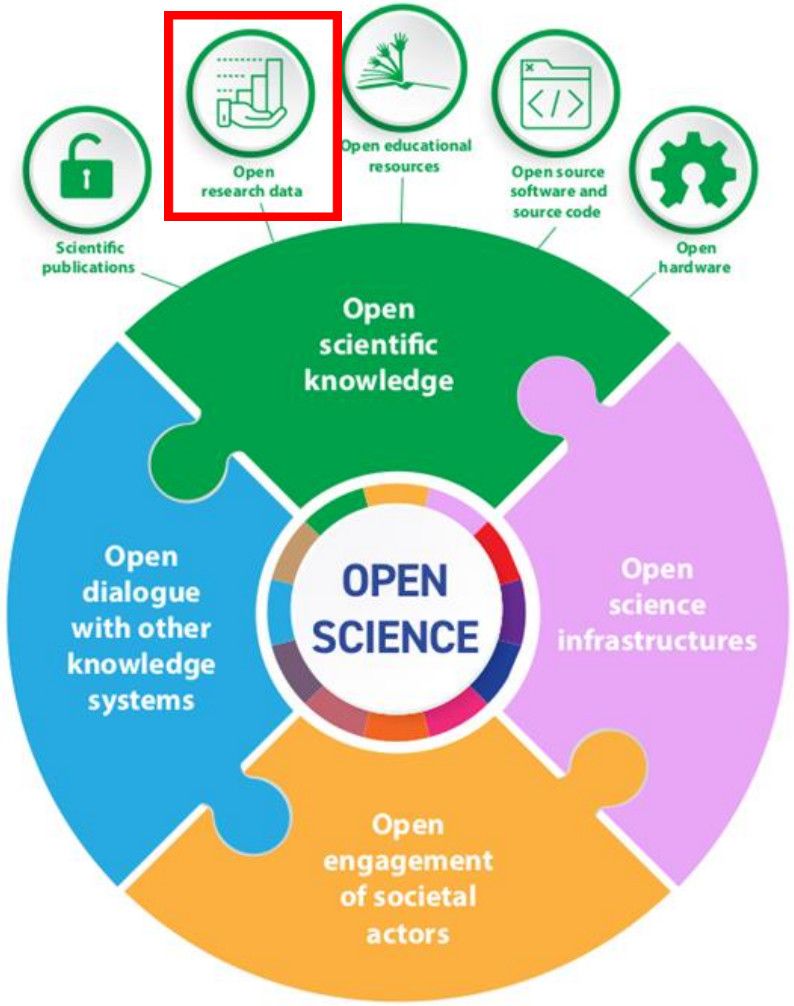
O que são dados de investigação?

São um **produto direto** ou **indireto** do processo de **investigação** e, como tal, necessários para validar os resultados científicos.

Existem **vários tipos** de dados:

- Dados numéricos
  - bases de dados, folhas de cálculo
- Dados textuais
  - documentos
- Dados de imagem, áudio, vídeo
- Suportes mistos
- Outros

Ciência Aberta e Dados Abertos



Fonte: https://unesdoc.unesco.org/ark:/48223/pf0000378841

Ciência Aberta

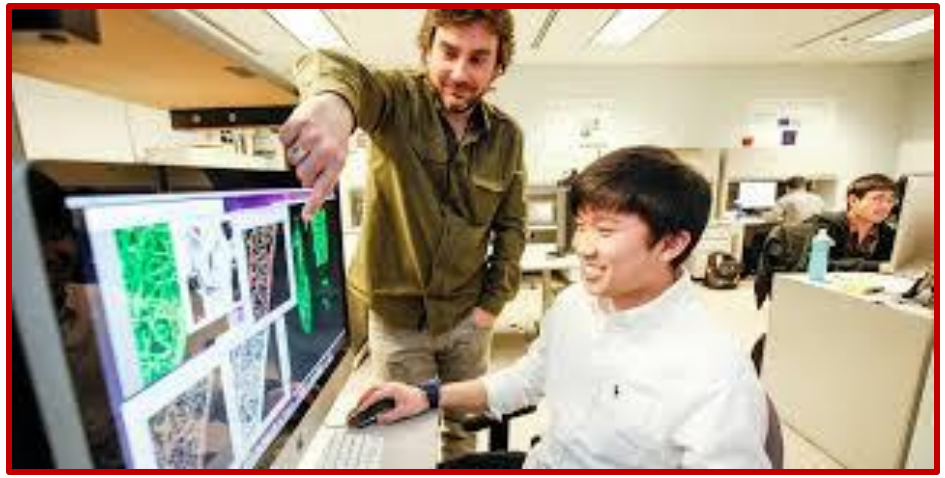
Prática de realizar investigação científica de forma **aberta**, **colaborativa** e **transparente**, em todos os domínios do conhecimento, colocando os resultados dessa investigação à disposição de todos.

Dados Abertos

Um dos pilares da Ciência Aberta.

Uma adequada gestão de dados contribui para...

- **Aumentar o impacto** da investigação;
- Tornar os **dados acessíveis**;
- **Validar os resultados** da investigação;
- **Preservar** os dados gerados;
- **Antecipar recursos**;
- Assegurar a **reutilização** dos dados.



Ciclo de vida dos dados de investigação

Ciclo de vida dos dados de investigação

Planeamento

- Primeiras reflexões quanto à produção, preservação e partilha dos dados;
- Antecipar recursos;
- Estabelecer as condições para que os dados da sua investigação tenham o maior impacto possível na ciência, mesmo após o fim do projeto.

Boa prática!

Elaboração de um **Plano de Gestão de Dados (PGD)**.

Ciclo de vida dos dados de investigação

Recolha

- Recolher informações sobre variáveis específicas de interesse;
- Garantir a qualidade dos dados;
- É também possível reutilizar dados existentes no seu projeto.

Boas práticas!

- **Documentação** e **metadados**
- Uso de **identificadores**
- **Armazenamento** dos dados
- **Dados sensíveis**
- **Proveniência** dos dados

Ciclo de vida dos dados de investigação

Processamento

- Conversão dos dados para um formato desejado;
- Preparação para análise.

Boas práticas!

- **Organização** dos dados
- **Proveniência** dos dados

Ciclo de vida dos dados de investigação

Análise

- Exploração dos dados recolhidos para começar a interpretar o conjunto de dados;
- Aplicação de modelos para identificar relações entre variáveis.
- De acordo com os diferentes tipos de dados (quantitativos ou qualitativos), os métodos de análise de dados serão diferentes.

Boas práticas!

- **Organização** dos dados
- **Armazenamento** e **back up** dos dados

Ciclo de vida dos dados de investigação

Preservação

- Preservar a longo prazo os dados que fundamentam os resultados da investigação e que têm valor;
- Curadoria e garantia de qualidade dos dados;
- Avaliar a reutilização dos dados gerados no decorrer do projeto.

Boas práticas!

- **Organização**
- Seleção de **local adequado** (usualmente um repositório)
- Uso de **identificadores**
- Atribuição de **licenças**
- Controlo de **acessos**

Ciclo de vida dos dados de investigação

Partilha

- Partilhar com parceiros no contexto do projeto de investigação;
- Partilhar com a comunidade de investigação e com a sociedade em geral;
- A partilha não implica dados abertos. Poderá ser estabelecido um acesso restrito.

Boas práticas!

- **Proteção de dados (RGPD)**
- **Dados sensíveis**
- Atribuição de **licenças**
- Controlo de **acessos**
- **Documentação** e **metadados**
- **Publicação** de dados

Ciclo de vida dos dados de investigação

Reutilização

- Utilizar os dados para outros fins que não aqueles para que foram originalmente recolhidos;
- Componente fundamental dos Princípios de Dados FAIR.

Boas práticas!

- **Descrição adequada**, com curadoria
- Termos e condições de partilha claros - **Licenças**
- **Integração** com outras fontes de dados
- **Proveniência** dos dados

Jornadas FCCN 2023

Planos de Gestão de Dados

João Cardoso

fct Fundação para a Ciência e a Tecnologia
FCCN

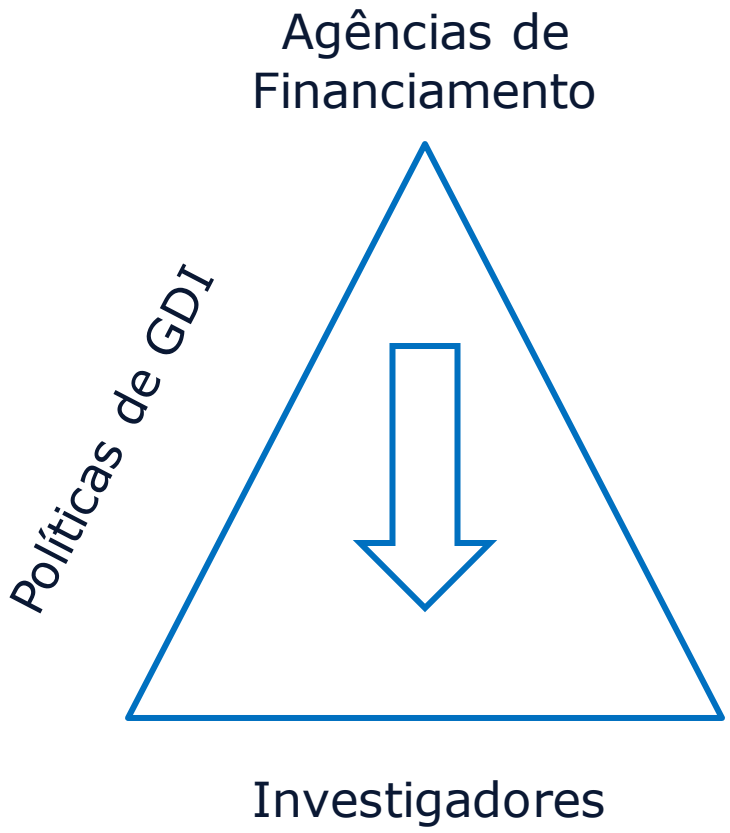
O que são Planos de Gestão de Dados?

Começa a ser consensual que quaisquer **políticas de gestão de dados de investigação** terão que ser aplicadas através de uma **abordagem top down**.

Sendo que o **conceito de PGD** surge como uma das **ferramentas** para a implementação desse tipo de abordagem.

Em particular com a ideia de encorajar a **submissão um PGD juntamente com candidaturas a financiamento.**

Esta **tendência tem continuado ao longo da última década**, com cada vez mais agências de financiamento a adotar essa prática.



O que são Planos de Gestão de Dados?

Um **PGD é um documento formal** que tem como objetivo fomentar o **planeamento e servir de referência durante a execução de atividades de gestão de dados.**

Em particular antecipando as **necessidades e requisitos ao longo de um projeto (de investigação)** numa determinada equipa, instituição ou organização.

Um PGD deve descrever **políticas** e **métodos** que são aplicados aos dados durante as múltiplas etapas do **ciclo de vida dos dados (de investigação).**

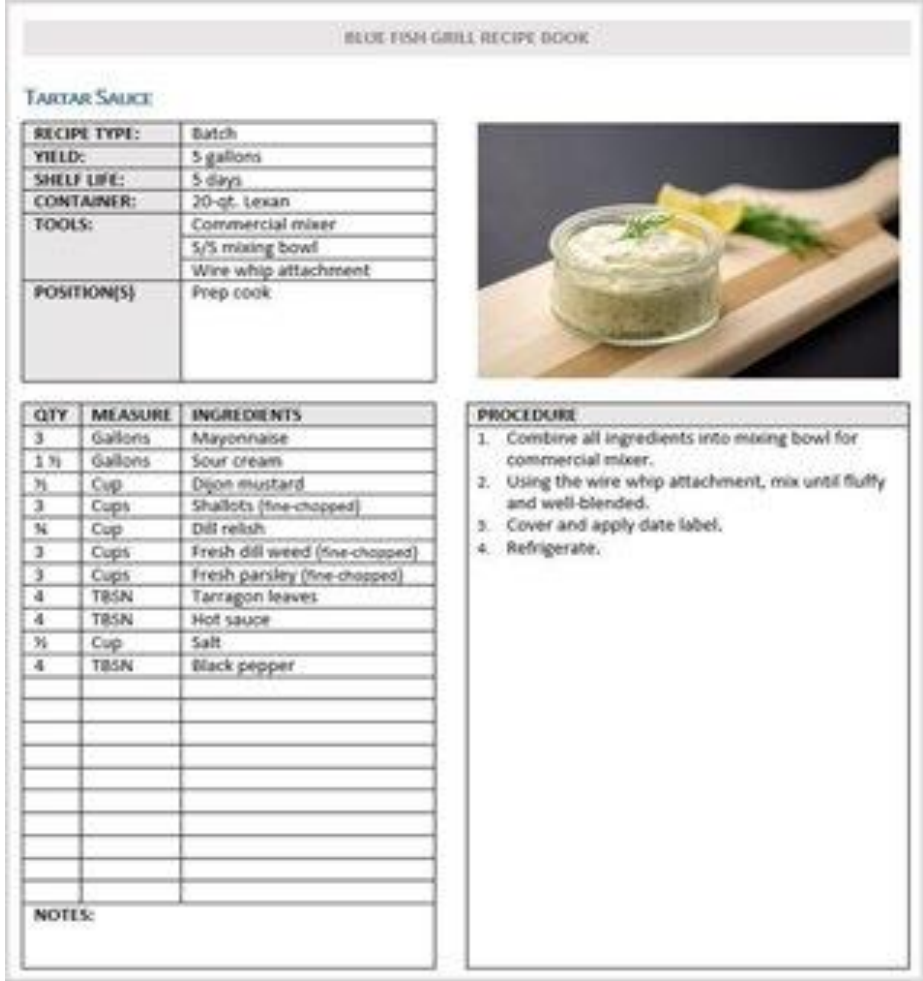
https://rdmkit.elixir-europe.org/data_life_cycle

O que são Planos de Gestão de Dados?

O PGD está para a gestão de dados de investigação como as fichas técnicas estão para a restauração.

Fichas técnicas => Descrição detalhada de como preparar receitas

PGD => Descrição detalhada de como gerir dados de investigação



BLUE FISH GRILL RECIPE BOOK

TARTAR SAUCE

RECIPE TYPE:	Batch
YIELD:	5 gallons
SHELF LIFE:	5 days
CONTAINER:	20-qt. Lexan
TOOLS:	Commercial mixer
	S/S mixing bowl
	Wire whip attachment
POSITION(S)	Prep cook

QTY	MEASURE	INGREDIENTS
3	Gallons	Mayonnaise
1 ½	Gallons	Sour cream
½	Cup	Dijon mustard
3	Cups	Shallots (fine-chopped)
¼	Cup	Dill relish
3	Cups	Fresh dill weed (fine-chopped)
3	Cups	Fresh parsley (fine-chopped)
4	TBSN	Tarragon leaves
4	TBSN	Hot sauce
½	Cup	Salt
4	TBSN	Black pepper

NOTES:

PROCEDURE

1. Combine all ingredients into mixing bowl for commercial mixer.
2. Using the wire whip attachment, mix until fluffy and well-blended.
3. Cover and apply date label.
4. Refrigerate.

O que são Planos de Gestão de Dados?

Via da proposta de valor:

- Os PGD são documentos que **agregam** em si **toda a informação relativa à política de GDI** de um determinado projeto (de investigação).
- São artefactos, são o **produto** e simultaneamente **estimulam o bom planeamento.**
- Permitem a **rastreabilidade** entre atividades e atores.
- Promovem o cumprimento dos **princípios FAIR**.
- Potenciam a **automação** de atividades de gestão de dados.
- Podem **acelerar** e **simplificar** os processos de investigação.

O que são Planos de Gestão de Dados?

Via da regulação:

Muitas agências de financiamento **encorajam a submissão de PGDs** juntamente com candidaturas a financiamento.

Esta política de **encorajamento** está em transição para **requerimento**.

Com os PGD as agências de financiamento procuram compreender **como irão ser cumpridos os princípios FAIR**.

Os **PGD devem que seguir templates** definidos pelas agências de financiamento.

Jornadas FCCN 2023

Ferramentas de elaboração de PGD

Filipa Pereira

fct Fundação para a Ciência e a Tecnologia
FCCN

Serviço - FCT|FCCN

Serviço a disponibilizar à comunidade

Sistema de PGD

https://bit.ly/36fr4aV

Suporte à seleção da ferramenta

Modelo FCT

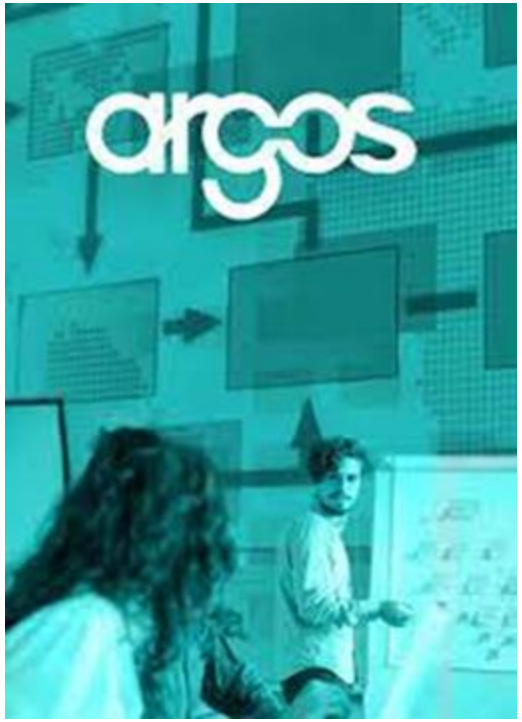
SCIENCE EUROPE

NWO

Existem várias ferramentas...

- DMPonline
  https://dmponline.dcc.ac.uk/
- ARGOS
  https://argos.openaire.eu/
- Data Stewardship Wizard
  https://ds-wizard.org/
- DMPTool
  https://dmptool.org/
- OpenDMP
  https://gitlab.eudat.eu/dmp/OpenAIRE-EUDAT-DMP-service-pilot
- ezDMP
  https://ezdmp.org
- RDMO - RDM Organiser
  https://rdmorganiser.github.io/en/
- easyDMP
  https://easydmp.eudat.eu
- DMP ARIADNEplus
  https://vast-lab.org/dmp/ariadneplus/form/

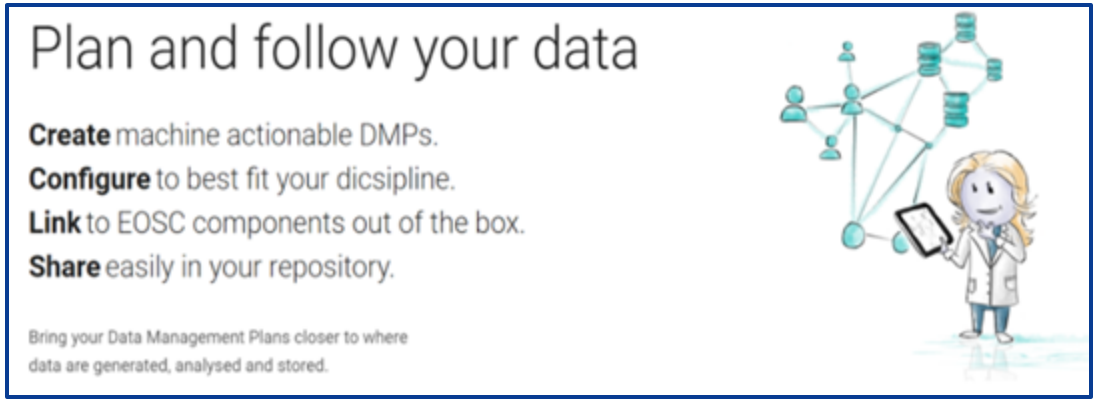
Solução



Serviço Cloud

https://argos.openaire.eu/

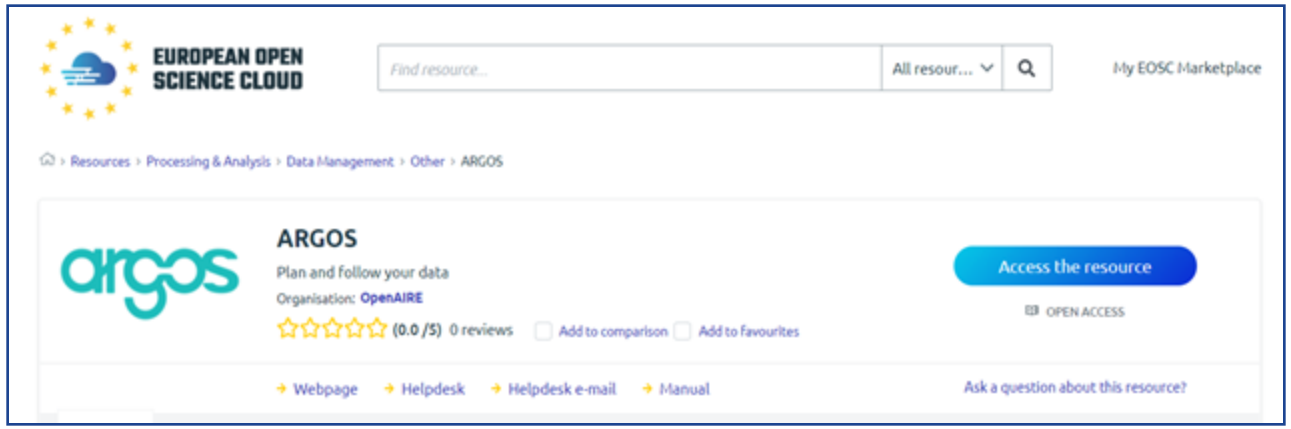
argos



- **Colaboração entre Infraestruturas Europeias** -> OpenAIRE EUDAT Collaborative Data Infrastructure Data shared and preserved across borders and disciplines
- **Projetos** -> Via OpenAIRE;
- **Flexível** -> Requisitos e necessidades do financiador;
- **Simples** -> Focada na interação com o utilizador;
- **FAIRness** dos dados;

argos

- **Adaptável** -> Diferentes disciplinas e domínios;
- **maDMPs** -> Criação de planos acionáveis por máquina;
- **Versão em Português**;
- Ligada à **European Open Science Cloud (EOSC)** -> Catálogo de serviços:



https://marketplace.eosc-portal.eu/services/argos

Estrutura

argos

PGD

- Pode ser exportado em vários formatos
- Composto por Datasets

- Nome, restrito ou público, instituição, investigador, financiador, *grant*, licença;
- Seleção do Dataset/Modelo de dados.

Dataset

- Conjunto de questões pré-definidas de acordo com modelo de instituição e/ou financiador

- Como serão tratados, armazenados e validados os dados gerados ou reutilizados durante o projeto de investigação.

Modelo PGD FCT

SCIENCE EUROPE

GUIDANCE FOR ORGANISATIONS:
CORE REQUIREMENTS FOR
DATA MANAGEMENT PLANS

GUIDANCE FOR RESEARCHERS:
Translating the Core Requirements into a DMP template
Guiding the Selection of Trustworthy Repositories

GUIDANCE FOR REVIEWERS:
Evaluation Rubric for Data Management Plans

https://scienceeurope.org/media/4brkxxe5/se_rdm_practical_guide_extended_final.pdf

Modelo PGD FCT - Tópicos

1 Informação sobre dados

2 Documentação e metadados

3 Armazenamento e segurança

4 Dados pessoais, IPR

5 Partilha e preservação

6 Responsabilidades e recursos

Jornadas FCCN 2023

Descrição do Exemplo

Projeto Quase

Projeto Quase

O Projeto Quase procura compreender os **fatores de risco associados a desastres de bicicleta**, no curso do seu uso em deslocações entre casa e os locais de trabalho.

Foram selecionados **100 voluntários** que terão montado na sua bicicleta um **dispositivo** que recolhe **dados de vídeo** e de **informação GPS**.

Na situação de o voluntário sofrer um **'quase acidente'** é pedido que **assinalem o evento pressionando um botão no dispositivo**, que procederá ao registo permanente dos dados referentes aos **5 minutos que antecederam o evento**.

(Inspirado no projeto *UCL 100 cyclist study* https://www.ucl.ac.uk/civil-environmental-geomatic-engineering/research-projects/2021/jun/100-cyclists-project)

Projeto Quase

A recolha de dados teve início no mês de **Janeiro de 2023** e decorrerá até ao final de **Janeiro de 2025**.

O dispositivo tem capacidade estimada para armazenar os dados referentes a **5 eventos**, sendo feita uma **sincronização automática com o servidor de dados da instituição** durante o carregamento do mesmo.

Ao serem depositados no servidor de dados da instituição, todos os dados são acompanhados de um **manual de operações que define as condições de recolha dos dados**.

O servidor de dados da instituição opera com uma **infraestrutura redundante**, sendo que os dados são **replicados em duas** instâncias do servidor de dados da instituição.

Projeto Quase

O Projeto Quase é um trabalho de investigação liderado pela **Escola Naval** sendo **financiado** pela FCT através do **grant FCCNPFA002**, com duração prevista de 2 anos.

Os **dados gerados** no âmbito do Projeto Quase são **propriedade da Escola Naval**, sendo que estarão disponíveis para consulta mediante **acesso controlado**.

Projeto Quase

A equipa de investigação é constituída por:

- João Cardoso - investigador principal
- Filipa Pereira – investigadora auxiliar

Exemplo - Projeto Quase

Serão considerados **cinco conjuntos de dados**:

Dados Espaciais

Imagens com dados de georreferenciação de origem externa

Formatos:
- .tif
- .tfw

Volume:
- ~1TB

Vídeo

Imagens vídeo do dispositivo (resolução 4k [UHD])

Formatos:
- .mp4

Volume:
- ~15GB

Dados de GPS

Imagens com dados de gps do dispositivo

Formatos:
- .gpx

Volume:
- < 1GB

Dados de Eventos

Ficheiro que associa evento e ficheiros de vídeo e dados GPS.

Formatos:
- .txt

Volume:
- < 1GB

Metadados:
- Marca temporal
- Identificador
- Ficheiro de vídeo
- Ficheiro de gps

Identificação de Ciclistas

Tabela que identifica os ciclistas participantes no projeto.

Formatos:
- .csv

Volume:
- < 1GB

Metadados:
- Nome do ciclista
- Idade do ciclista
- Número do cartão de cidadão
- Identificador

Exemplo - Projeto Quase

Serão considerados **cinco conjuntos de dados**:

Dados Espaciais

Imagens com dados de georreferenciação de origem externa

Formatos:
- .tif
- .tfw

Volume:
- ~1TB

Vídeo

Imagens vídeo do dispositivo (resolução 4k [UHD])

Formatos:
- .mp4

Volume:
- ~15GB

Dados de GPS

Imagens com dados de gps do dispositivo

Formatos:
- .gpx

Volume:
- < 1GB

Dados de Eventos

Ficheiro que associa evento e ficheiros de vídeo e dados GPS.

Formatos:
- .txt

Volume:
- < 1GB

Metadados:
- Marca temporal
- Identificador
- Ficheiro de vídeo
- Ficheiro de gps

Identificação de Ciclistas

Tabela que identifica os ciclistas participantes no projeto.

Formatos:
- .csv

Volume:
- < 1GB

Metadados:
- Nome do ciclista
- Idade do ciclista
- Número do cartão de cidadão
- Identificador

Demonstração de utilização do Argos!

argos

https://argos.openaire.eu/

Projeto Quase

1 Informação sobre dados

Dados Espaciais

Imagens com dados de georreferenciação de origem externa

Formatos:
- .tif
- .tfw

Volume:
- ~1TB

Vídeo

Imagens vídeo do dispositivo (resolução 4k [UHD])

Formatos:
- .mp4

Volume:
- ~15GB

Dados de GPS

Imagens com dados de gps do dispositivo

Formatos:
- .gpx

Volume:
- < 1GB

Dados de Eventos

Ficheiro que associa evento e ficheiros de vídeo e dados GPS.

Formatos:
- .txt

Volume:
- < 1GB

Metadados:
- Marca temporal
- Identificador
- Ficheiro de vídeo
- Ficheiro de gps

Identificação de Ciclistas

Tabela que identifica os ciclistas participantes no projeto.

Formatos:
- .csv

Volume:
- < 1GB

Metadados:
- Nome do ciclista
- Idade do ciclista
- Número do cartão de cidadão
- Identificador

Projeto Quase

2 Documentação e metadados

O dispositivo tem capacidade estimada para armazenar os dados referentes a 5 eventos, sendo feita uma **sincronização automática com o servidor de dados da instituição** durante o carregamento do mesmo.

Ao serem depositados no servidor de dados da instituição, todos os dados são acompanhados de um **manual de operações que define as condições de recolha dos dados**.

O servidor de dados da instituição opera com uma **infraestrutura redundante**, sendo que os dados são **replicados em duas** instâncias do servidor de dados da instituição.

Projeto Quase

2 Documentação e metadados

Dados Espaciais

Imagens com dados de georreferenciação de origem externa

Formatos:
- .tif
- .tfw

Volume:
- ~1TB

Vídeo

Imagens vídeo do dispositivo (resolução 4k [UHD])

Formatos:
- .mp4

Volume:
- ~15GB

Dados de GPS

Imagens com dados de gps do dispositivo

Formatos:
- .gpx

Volume:
- < 1GB

Dados de Eventos

Ficheiro que associa evento e ficheiros de vídeo e dados GPS.

Formatos:
- .txt

Volume:
- < 1GB

Metadados:
- Marca temporal
- Identificador
- Ficheiro de vídeo
- Ficheiro de gps

Identificação de Ciclistas

Tabela que identifica os ciclistas participantes no projeto.

Formatos:
- .csv

Volume:
- < 1GB

Metadados:
- Nome do ciclista
- Idade do ciclista
- Número do cartão de cidadão
- Identificador

Projeto Quase

3 Armazenamento e segurança

O dispositivo tem capacidade estimada para armazenar os dados referentes a 5 eventos, sendo feita uma **sincronização automática com o servidor de dados da instituição** durante o carregamento do mesmo.

Ao serem depositados no servidor de dados da instituição, todos os dados são acompanhados de um **manual de operações que define as condições de recolha dos dados**.

O servidor de dados da instituição opera com uma **infraestrutura redundante**, sendo que os dados são **replicados em duas** instâncias do servidor de dados da instituição.

Projeto Quase

4 Dados pessoais, IPR e ownership

Dados Espaciais

Imagens com dados de georreferenciação de origem externa

Formatos:
- .tif
- .tfw

Volume:
- ~1TB

Vídeo

Imagens vídeo do dispositivo (resolução 4k [UHD])

Formatos:
- .mp4

Volume:
- ~15GB

Dados de GPS

Imagens com dados de gps do dispositivo

Formatos:
- .gpx

Volume:
- < 1GB

Dados de Eventos

Ficheiro que associa evento e ficheiros de vídeo e dados GPS.

Formatos:
- .txt

Volume:
- < 1GB

Metadados:
- Marca temporal
- Identificador
- Ficheiro de vídeo
- Ficheiro de gps

Identificação de Ciclistas

Tabela que identifica os ciclistas participantes no projeto.

Formatos:
- .csv

Volume:
- < 1GB

Metadados:
- Nome do ciclista
- Idade do ciclista
- Número do cartão de cidadão
- Identificador

Projeto Quase

4 Dados pessoais, IPR e ownership

O Projeto Quase é um trabalho de investigação liderado pela **Escola Naval** sendo **financiado** pela FCT através do **grant FCCNPFA001**, com duração prevista de 2 anos.

Os **dados gerados** no âmbito do Projeto Quase são **propriedade da Escola Naval**. Sendo que estarão disponíveis para consulta mediante **acesso controlado**.

Projeto Quase

5 Partilha e preservação

- Serviço de Repositório de Dados de Investigação
- Baseado no software Dataverse
- Sistema de PID (DOI) suportado pelo DataCite
- Licenças por defeito Creative Commons 4.0

Projeto Quase

6 Responsabilidades e recursos

Equipa de investigação constituída por:

João Cardoso - investigador principal

Recolha dos dados; produção de metadados; qualidade dos dados; partilha dos dados; implementação e respetiva reavaliação; preparação de PGD.

Filipa Pereira – investigadora auxiliar

Recolha dos dados; Produção de metadados; Qualidade dos dados; Data protection officer; Armazenamento e cópia de segurança; Arquivo dos dados; Partilha dos dados; Atualização de PGD.

Projeto Quase

6 Responsabilidades e recursos

- Acesso ao repositório de dados de investigação (avaliar se existe algum custo associado);
- Dispositivos de hardware;
- Software;
- Serviços cloud.

Questões?

Jornadas FCCN 2023

Serviço de Repositório de Dados

João Cardoso

Porquê depositar/preservar dados de investigação?

- Manter e preservar de forma **fidedigna** e **confiável** os dados de investigação;
- Processo essencial para o **acesso livre** e **sustentável a longo prazo**;
- Recomenda-se a seleção de um repositório que seja considerado **confiável**.

Partilha e preservação a longo prazo

Decisões importantes a tomar:

- Que dados serão preservados e por quanto tempo?
- Que dados estarão disponíveis para reutilização?
- Quando é que os dados estarão disponíveis para reutilização?
- Por quanto tempo estarão os dados disponíveis?

Repositório Polen

Servir os **beneficiários de instrumentos de financiamento da FCT**, nomeadamente a comunidade de investigação e de ensino superior.

Dar resposta às seguintes necessidades:

- Promover as **boas práticas** e os **princípios da Ciência Aberta**;
- Assegurar o cumprimento dos **Princípios de Dados FAIR**;
- Apoiar o **depósito**, **preservação** e **disseminação** de dados de investigação resultantes de fundos públicos;

Repositório Polen

- Dar suporte à **Política de gestão e partilha de dados da FCT**;
- Promover a adotação do **quadro normativo do PTCRIS**;
- Integrar com o **RCAAP**;
- Alinhar com a parceria **EOSC**.

Dataverse

Software desenvolvido pelo I**nstitute for Quantitative Social Science (IQSS)**, da Universidade de Harvard.

Software **open source**.

Gestão e partilha de dados de investigação:

- Processo de depósito simplificado;
- Citação de dados;
- Controlo de acesso aos dados;
- *Standards* de metadados;
- Políticas de partilha de dados.

Instâncias Dataverse a nível global

Instâncias Dataverse a nível nacional

Universidade do Minho

https://datarepositorium.uminho.pt/

ipb INSTITUTO POLITÉCNICO DE BRAGANÇA

https://dados.ipb.pt/

universidade de aveiro

https://dunas.ua.pt/

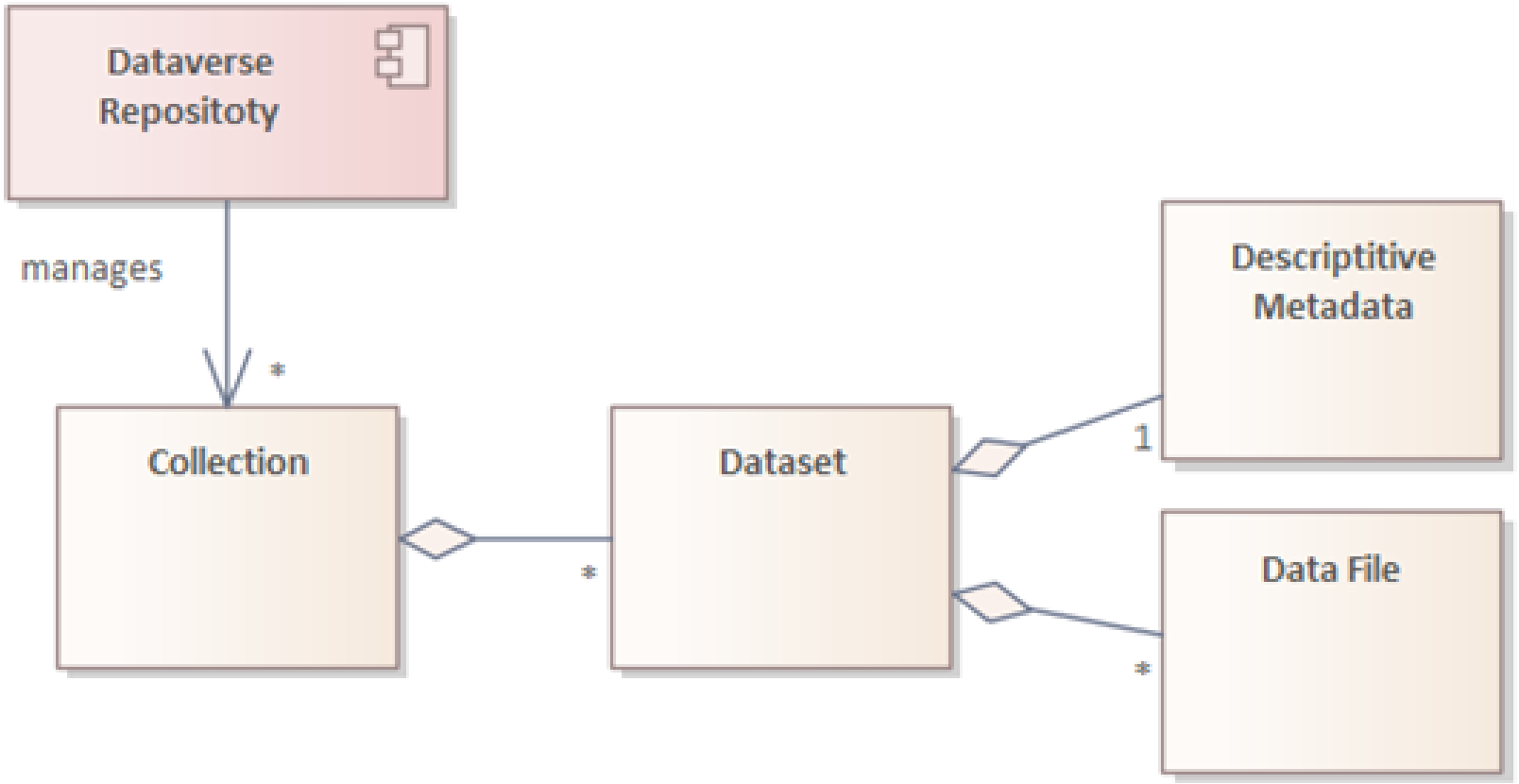
BioData.pt

https://dmportal.biodata.pt/

polen repositório ***Fase Piloto***

https://repositorio.polen.fccn.pt/

Dataverse - Estrutura



Demonstração de utilização do Repositório Polen!

MIA
Questões?

Jornadas FCCN 2023

Muito obrigado pela atenção!

Filipa Pereira 8A13-4EA4-A512

João Cardoso 2012-B202-C893

helpdesk.polen@fccn.pt

Jornadas FCCN 2023

pdi.n GESTÃO DE DADOS DE INVESTIGAÇÃO

Sessão Dados de Investigação

27 de junho | 16h30 |Sala 4

fct Fundação para a Ciência e a Tecnologia FCCN

Patrocinadores

Platina

EBSCO
ELSEVIER
Microsoft
warpcom
cisco
FORTINET
axians
officelan
SECURNET
CHECK POINT
SYSCRUM
PAESSLER
THE MONITORING EXPERTS
ORACLE
NVIDIA
paloalto
software one

Ouro

ACS Publications
Most Trusted. Most Cited. Most Read.
Clarivate
CAMBRIDGE UNIVERSITY PRESS
DE GRUYTER
HUAWEI
DIVULTEC
LOGICALIS
Architects of Change
Sage
SPRINGER NATURE
tp-link
wavecom
Bravantic
itcenter
reallife

Prata

ROYAL SOCIETY OF CHEMISTRY
IOP Publishing
MEO EMPRESAS
aws
emerald PUBLISHING
IEEE
Logiser

Organização

fct Fundação para a Ciência e a Tecnologia
FCCN