FÓRUM
GESTÃO DE DADOS
DE INVESTIGAÇÃO

Folha informativa 2022
https://forumgdi.rcaap.pt
DOI: https://doi.org/10.5281/zenodo.7082551

Como preparar dados para depositar e publicar no repositório

Dicas práticas para Investigadores, curadores de dados,
gestores de laboratório e responsáveis de unidades de investigação

Objetivo: Dotar os investigadores, gestores de dados e públicos relevantes, de ferramentas que promovam o depósito de dados, cumprindo determinados procedimentos, indo de encontro às orientações emanadas pelos organismos financiadores (ex. UE) e instituições nacionais e internacionais especializadas na GDI, no contexto da Ciência Aberta e na implementação dos princípios FAIR.

Formato dos dados

No início de um projeto existe a necessidade de criação de um **Plano de Gestão de Dados**, requisito de muito projetos financiados (ex. FCT e CE), sendo que uma das primeiras decisões a tomar é definir quais os formatos de ficheiros a utilizar, uma vez que existem **formatos padrão e formatos proprietários**. Estes últimos podem acarretar consequências negativas, tornando os dados menos interoperáveis.
O grande foco será sempre a sua **preservação**, o **acesso** continuado aos dados, a sua **interoperabilidade** e, tanto quanto possível, o cumprimento dos **princípios FAIR**. Não existem garantias se formatos de ficheiros proprietários usados diariamente existirão no futuro. Por exemplo, formatos de ficheiros da Microsoft (Word e Excel), podem estar em uso agora, mas isto não invalida a possibilidade de no futuro se tornarem obsoletos.

Aspeto a reter

Devemos referir que tornar os **dados FAIR** é cada vez mais uma responsabilidade conjunta dos investigadores e dos repositórios. Neste contexto e por forma a otimizar o arquivo, disseminação e reutilização de dados, o desenvolvimento de metadados ricos é um requisito dos **princípios FAIR** (Findable, Accessible, Interoperable and Reusable).

Os **formatos dos ficheiros** podem apresentar-se com ou sem perdas de informação: ou seja, se os dados são ou não comprimidos.
A compressão de ficheiros tem como objetivo a redução do tamanho dos mesmos por intermédio da remoção de informação redundante. Atualmente, podemos considerar os formatos abaixo indicados como apropriados e aceitáveis para preservação a longo prazo, assegurando assim o acesso continuado.

Tipos de dados	Apropriado	Aceitável	Não apropriado
Dados tabulares com metadados extensos	.csv - .hdf5	.txt - .html - .tex - .por	
Dados tabulares com metadados mínimos	.csv - .tab - .ods - SQL	.xml if appropriate DTD - .xlsx	.xls - .xlsb
Dados textuais	pdf. - .txt - .odt - .odm - .tex - .md - .htm - .xml	.pptx - PDF with embedded forms - .rtf	.doc - .ppt
Código	.m - .R - .py - .iypnb - rstudio - .rmd - NetCDF	.sdd	.mat - rdata
Dados de imagem digital	.tif - .png - .svg - .jpeg	.jpg - .jp2 - .tif - .tiff - .pdf - GIF - BMP	.indd - .ait - .psd
Dados de áudio digital	.flac - .wav - .ogg	.mp3 - .mp4 - .aif	
Dados de vídeo digital	.mp4 - .mj2 - .avi - .mkv	.txt - .html - .tex - .por	.wmv - .mov
Dados geoespaciais	NetCDF, tabular GIS Atribute data, shp - .shx .dbf - .prj - .sbx - .sbn - PostGIS - .tif - .tfw - GeoJSON	.ogm - .webm	
Dados vetorais e matriciais	.dwg - .dxf - .x3d - .x3dv - .x3db - .pdf - .PDF3D		
Dados genéricos	.xml - .json - .rdf		

Fonte: MOOC "O essencial da gestão de dados de investigação" em https://www.nau.edu.pt/pt/curso/o-essencial-da-gestao-de-dados-de-investigacao/

Metadados e vocabulários controlados

Metadados

A **caracterização dos conjuntos de dados** com recurso à utilização de metadados e vocabulários adequados é fundamental, para a sua correta descrição, seguindo as boas práticas no que respeita à partilha dos dados de investigação. A descrição tem impacto na forma como os dados podem ser encontrados e compreendidos por outros investigadores e, assim, corretamente agregados pelas infraestruturas de informação científica de referência, como é o caso do **OpenAIRE.**

Existem esquemas de **metadados disciplinares** que permitem uma descrição mais detalhada e que respondem às especificidades de cada área de investigação.
Pode consultar o **diretório de esquemas de metadados da RDA** em:
https://rdamsc.bath.ac.uk/

Quanto aos metadados podemos identificar 3 grandes categorias:

- **Descritivos**: Título, autor, resumo, palavras-chave: que auxiliam os utilizadores a descobrir os recursos online, através da pesquisa em bases de dados e motores de busca;
- **Administrativos**: Dados para preservação, direitos de autor, licenças, termos de uso, embargo e dados técnicos sobre os formatos;
- **Estruturais**: Informação de como diferentes componentes de um conjunto de dados associados, se relacionam entre si.

Vocabulários

Quando falamos em vocabulários controlados...

... falamos de uma linguagem composta por termos que podem ser organizados e estruturados de forma relacional ou alfabética. Trata-se basicamente de uma lista (hierárquica ou não) de termos a serem utilizados no processo de indexação - ou representação temática - de um documento com (meta)dados.
Para se utilizarem **vocabulários adequados** no processo de depósito dos dados, algumas iniciativas disponibilizam conjuntos de vocabulários controlados, como é o caso do **DDI** (**Data Documentation Initiative**), cujos vocabulários podem ser utilizados para descrever os conjuntos de dados, ao longo do seu ciclo de vida, nas áreas das ciências sociais, comportamentais, económicas, e da saúde.

Alguns exemplos

Outro exemplo para a área das ciências sociais é o **CESSDA Vocabulary Service** (baseado no **DDI**), uma ferramenta que permite aos utilizadores procurar, pesquisar e fazer o download de vocabulários controlados numa grande variedade de idiomas. Esta ferramenta é útil para investigadores, gestores e administradores de dados, tradutores de vocabulários controlados, e ainda para repositórios/arquivos que partilhem metadados.
Outras iniciativas que disponibilizam **vocabulários controlados para várias áreas disciplinares** podem ser consultadas em

https://guides.lib.utexas.edu/metadata-basics/controlled-vocabs

Dados Sensíveis

O que são?

Entende-se por **dados pessoais/sensíveis**, a informação que possibilite a identificação de forma direta ou indireta do titular dos mesmos. Devem ser tomadas medidas contra a divulgação indesejada deste tipo de informação. A aplicação pode ser exigida por razões legais, éticas, por questões relacionadas com a privacidade do indivíduo, ou por questões de propriedade.

Regulamentação

A UE tem regulamentação no que toca à salvaguarda de dados pessoais (como é o caso do **Regulamento Geral de Protecção de Dados** - RGPD) e a dados sensíveis não pessoais.

https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=OJ:L:2016:119:FULL&from=EN

São dados sensíveis podem ser:

- dados pessoais: tais como nomes ou números de identificação, dados biométricos, mentais, económicos, culturais ou sociais, incluindo também dados de localização de GPS ou de telemóveis;
- dados confidenciais: segredos comerciais, investigações, dados protegidos por direitos de propriedade intelectual;
- dados de segurança: palavras-passe, informação financeira, segurança nacional, informação militar, entre outros;
- uma combinação de diferentes conjuntos de dados que podem resultar na obtenção de dados sensíveis ou pessoais;
- dados biológicos: espécies ameaçadas (vegetais ou animais), onde a sua sobrevivência depende da protecção dos seus dados de localização (comunidade de biodiversidade);
- metadados pessoais e sensíveis.

Não esquecer: no processo de tratamento de dados sensíveis, deve ser dada especial atenção à recolha, processamento, tratamento e armazenamento de dados ao longo de todo o processo de investigação. Em particular, os dados de investigação que contenham dados pessoais com os quais uma pessoa possa, direta ou indiretamente, ser identificada. Isto diz respeito tanto aos dados textuais como aos dados de imagem e som. No caso de dados sensíveis, deve ser obtido o consentimento plenamente informado para a recolha, processamento, armazenamento e eventual publicação dos dados.

Nota metodológica (*README file*)

O que é e porquê?

No momento de depósito dos conjuntos de dados é imprescindível a integração de uma nota metodológica, comumente conhecida como ficheiro **README** (ou README File). Este ficheiro deve ser criado no **formato .txt** e deverá conter toda a **informação necessária para garantir que os dados possam ser corretamente interpretados**, ou seja, as linhas orientadoras para que qualquer pessoa possa **reutilizar os dados e/ou replicar a metodologia aplicada**.

A identificação de todo o **processo de recolha dos dados, instrumentos/software de recolha, intervalo de tempo de recolha dos dados, forma de tratamento (instrumentos utilizados) e metodologia de análise**, são parte da informação que idealmente deverá ser integrada. Aqui não podemos esquecer a informação relativa às licenças de utilização/reutilização dos dados. Só assim, estamos a ser transparentes e a cumprir com os fundamentos da Ciência Aberta.

Consulte estes exemplos de modelos de *README file* produzido pela **Cornell University** e pelo **consórcio dataverse de instituições da Noruega**, respetivamente:

☞ https://cornell.app.box.com/v/ReadmeTemplate

☞ https://site.uit.no/dataverseno/deposit/prepare/#readmefile

Recursos úteis

Formatos de dados

DANS (the Dutch national centre of expertise and repository for research data)

https://dans.knaw.nl/en/file-formats/

UK Data Service

https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

Ferramentas de anonimização de dados

Amnesia
https://amnesia.openaire.eu/

Datprof
https://www.datprof.com/products/datprof-privacy/

ARX
https://arx.deidentifier.org/overview/

Anonimatron
https://realrolfje.github.io/anonimatron/

Argus
https://qosient.com/argus/anonymization.shtml

Metadados/ vocabulários controlados

CESSDA Vocabulary Service
https://vocabularies.cessda.eu/#!discover

Vocabulários oceanográficos
https://www.bodc.ac.uk/resources/vocabularies/

Data Documentation Initiative (DDI)
https://www.ddialliance.org/

Diretório de esquemas de metadados gerido pela Research Data Alliance
https://rdamsc.bath.ac.uk

FAIR

Princípios FAIR
https://www.go-fair.org/fair-principles/

How FAIR are your data?
Checklist de Jones e Grootveld (2017)
https://www.cessda.eu/content/download/3845/35038/file/20170707_How_FAIR_are_your_data_Jones.pdf

Artigo "The FAIR Guiding Principles for scientific data management and stewardship"
https://www.nature.com/articles/sdata201618

Factsheet: The FAIR Data Principles
https://www.openaire.eu/factsheet-fair-data-principles/view-document

Alguns guias do OpenAIRE:

Data formats for preservation: What you need to know when creating a DMP
https://www.openaire.eu/data-formats-preservation-guide

How to deal with non-digital data: The benefits of digitising data
https://www.openaire.eu/non-digital-data-guide

How to deal with sensitive data: Learn how to preserve your sensitive data safely.
https://www.openaire.eu/sensitive-data-guide

Raw data, backup and versioning: What you need to know to preserve your research data
https://www.openaire.eu/RAW-DATA-BACKUP-AND-VERSIONING

Referências importantes:

Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles
https://datascience.codata.org/articles/10.5334/dsj-2021-012/

How can e-infrastructures deal with the sensitive data challenge (Working Paper)
https://b2share.eudat.eu/records/3d1dfb9b889c4022ae7b308df009fcc9

Deposit you data in a data repository for long-term preservation
https://bit.ly/coggle_DepositData_DataRepository

Guia: Gestão de dados de investigação (ISCTE)
https://bibliosubject.iscte-iul.pt/sp4/subjects/guide.php?subject=GDI

Fórum de Gestão de Dados de Investigação

Grupo de Trabalho Repositórios de Dados: Tecnologia, organização e certificação

website https://forumgdi.rcaap.pt/grupos-de-trabalho/

Outros recursos deste grupo já disponíveis
https://forumgdi.rcaap.pt/grupos-de-trabalho/gt-repositorios-de-dados/gt-repositorios-resultados/

Recursos do GT da Formação e competências para Gestão e Dados FAIR
http://forumgdi.rcaap.pt/grupos-de-trabalho/gt-formacao/

Créditos
Gravuras e icons - Designed by Eightonesix, Stories, Vectorjuice and Dooder / Freepik