10 guias, manuais digitais, perguntas frequentes, e boas práticas

Centro de Gestão de Dados de Investigação Interdisciplinar da Universidade de Lisboa (iRe:Search)

Dezembro 2025
V1.0

PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU

Índice

Introdução 5
1. Ciência Aberta e Dados de Investigação Abertos 6
1.1. Conceitos Chave 6
1.2. Introdução à Ciência Aberta e Dados Abertos 7
1.3. Estratégia Institucional e Contexto Legal 12
1.4. Caso de Estudo: Implementação dos Princípios FAIR 14
1.5. Sistemas e Repositórios de Ciência Aberta 15
1.6. Obrigações e Boas Práticas 16
2. Gestão e Preservação de Dados Abertos, Condicionalismos Legais e Regulamentares 18
2.1. Conceitos Chave 18
2.2. Melhores Práticas em Gestão de Dados a Longo Prazo 19
2.3. Estratégias de Armazenamento e Preservação 20
2.4. Conformidade Legal e Regulamentar 21
2.5. Caso de Estudo: Desafios de Preservação (Smart City) 23
2.6. Obrigações e Boas Práticas 23
3. Direitos de Autor e Direitos Conexos 25
3.1. Conceitos Chave 25
3.2. Fundamentos Legais e Aplicação 25
3.3. Outras Medidas de Proteção e Justificações para Restrição 26
3.4. Declaração de Disponibilidade de Dados (DAS) 27
3.5. Modelos de Acesso Aberto (AA) para Publicações 27
3.6. Opções de Licenciamento Aberto 28
3.7. Mandatos de Licenciamento e Retenção de Direitos 30
3.8. Exemplo da Licença MIT 30
3.9. Caso de Estudo: Escolha de Licença (Smart City) 31
3.10. Obrigações e Boas Práticas 33
4. Regulamento Geral de Proteção de Dados (RGPD) 35
4.1. Conceitos Chave 35
4.2. Aplicação do RGPD na Gestão de Dados Pessoais - Legislação Aplicável e Enquadramento Nacional 36
4.3. Bases Legais, Transparência e PGD 37

4.4. Proteção de Dados desde a Conceção 38
4.5. Avaliação de Impacto sobre a Proteção de Dados (DPIA) 39
4.6. Ciclo de Vida dos Dados Pessoais e Salvaguardas 40
4.7. Caso de Estudo: Anonimização e Consentimento 42
4.8. Obrigações e Boas Práticas 43
5. Normas Internacionais e Boas Práticas 45
5.1. Conceitos Chave 45
5.2. Standards Internacionais para a Gestão de Dados 45
5.3. Frameworks e Boas Práticas na Ciência Aberta 49
5.4. Casos de Estudo (Exemplos de Aplicação de Standards) 52
5.5. Obrigações e Boas Práticas 52
6. Ferramentas e Recursos Disponíveis 54
6.1. Conceitos Chave 54
6.2. Repositórios e Plataformas para Gestão de Dados 55
6.3. Caso de Estudo: Fluxo de Trabalho com Repositório 61
6.4. Obrigações e Boas Práticas 62
7. Ética e Transparência 63
7.1. Conceitos Chave 63
7.2. Considerações Éticas na Gestão e Partilha de Dados 64
7.3. Integridade Científica e Protocolos Éticos 67
7.4. Casos de Estudo: Dilemas Éticos na Partilha 70
7.5. Obrigações e Boas Práticas 72
8. Reprodutibilidade 73
8.1. Conceitos Chave 73
8.2. O Papel dos Dados na Reprodutibilidade Científica 74
8.3. Disponibilização de Dados e Resultados para Validação 77
8.4. Casos de Estudo: Replicação e Reprodução de Resultados 83
8.5. Obrigações e Boas Práticas 84
9. Princípios de Gestão de Dados de Investigação e DMPs 89
9.1. Conceitos Chave 89
9.2. Princípios Fundamentais do RDM e Requisitos para PGDs 89
9.3. Estrutura e Ferramentas para PGDs 97
9.4. Casos de Estudo: PGDs Reais de Projetos Europeus 98
9.5. Obrigações e Boas Práticas 100

9.6. Modelos de PGD para a ULisboa 101
9.6.1. Modelo de PDG completo 101
9.7. Modelo de PDG simplificado 105
10. Utilização de Inteligência Artificial (IA) em gestão de dados e investigação 108
10.1. Conceitos Chave 108
10.2. IA no Ciclo de Investigação e Implicações Éticas/Legais 109
10.3. O Enquadramento Legal Europeu (AI Act) 116
10.4. Privacidade, Conformidade (RGPD) e Integridade com a IA 117
10.5. Casos de Estudo 119
10.6. Obrigações e Boas Práticas 120
Anexo I – Indicadores de sucesso sugeridos 122
Anexo II – Sumário de obrigações e boas práticas 126

Introdução

Este documento inclui os conteúdos detalhados para 10 guias e manuais digitais, cobrindo os tópicos específicos identificados na Tarefa A4.2: Ciência Aberta e Dados de Investigação Abertos, Gestão e Preservação de Dados, Direitos de Autor, RGPD, Normas Internacionais, Ferramentas e Recursos, Ética e Transparência, Reprodutibilidade, Princípios de Gestão de Dados de Investigação e DMPs, e Utilização de Inteligência Artificial. Para além de servirem de base para a realização das formações na A4, este documento apresenta guias, autónomos, sobre cada um destes temas, com recomendações e boas práticas. É complementado pelo DL3.1, o catálogo de recursos online, com resumos explicativos de cada recurso disponível e perguntas frequentes (documento anexo ao DL3.1), e pelos conteúdos formativos da A4 que aprofundam os temas destes guias, incluindo mais detalhes sobre os casos de estudo, exemplos e exercícios.

Os entregáveis desta tarefa A3 seguem os princípios de reutilizar, referir, e resumir materiais e ferramentas existentes e disponíveis para uso aberto por parte de investigadores. Isto inclui por exemplo, documentação e recursos das Universidades analisadas na Atividade 1, e ferramentas como as identificadas no DL3.1. O foco da tarefa A3 é na identificação de recursos reutilizáveis ou que possam servir de inspiração para a criação de novos materiais, promovendo a eficiência e a evitando a duplicação de esforços. Isto potencia a sustentabilidade, integrando recursos técnicos, normativos e comunitários já existentes, para benefício da comunidade e acelerando a adoção de boas práticas de Ciência Aberta.

1. Ciência Aberta e Dados de Investigação Abertos

1.1. Conceitos Chave

A Ciência Aberta (CA) é um conceito inclusivo que visa a co-criação de conhecimento científico por e para toda a humanidade. A CA é uma modalidade participativa e colaborativa de fazer pesquisa cujo objetivo é aumentar a transparência, a confiabilidade, a inclusividade e a reprodutibilidade da pesquisa. Este movimento estratégico procura tornar todo o processo de investigação (métodos, dados, publicações, software) acessível a todos, promovendo o acesso livre e transparente para reforçar a colaboração, a reprodutibilidade e o impacto social da ciência. O objetivo da CA é acelerar a descoberta e melhorar a qualidade da investigação, e envolve a participação de todos os atores do conhecimento relevantes.

O Princípio Orientador da CA é fundamental para orientar as políticas institucionais: "**Tão aberto quanto possível, tão fechado quanto necessário**".

Os pilares-chave da CA, alinhados com a UNESCO, incluem o conhecimento científico aberto (publicações, dados FAIR, software), a infraestrutura científica aberta, a comunicação científica, o envolvimento aberto de atores sociais e o diálogo aberto com outros sistemas de conhecimento. Os componentes da CA incluem também: colaboração aberta; utilização de metodologias abertas (e.g., open notebook); pré-registo das experiências; open peer review; e recursos educacionais abertos, bem como pre-registration e registered reports.

A Gestão de Dados de Investigação (GDI/RDM) é a prática de apoio indispensável para assegurar a qualidade e a reutilização dos dados, sendo orientada pelos Princípios FAIR. A GDI inclui a recolha, organização, curadoria, armazenamento e documentação dos dados, garantindo a segurança, controlo de qualidade, alocação de PIDs, licenças e regras para partilha

O conceito de Dados de Investigação (Research Data) deve ser abrangente, incluindo dados em bruto e processados, dados de simulação e observação, transcrições de inquéritos, código, algoritmos, scripts, dados audiovisuais e objetos físicos (amostras, artefactos). Outros resultados incluem software, algoritmos, protocolos, modelos, workflows e electronic notebooks.

O foco de vários financiadores de investigação (p.ex. EU/Horizonte Europa, FCT) nestes

conceitos visa assegurar a transparência, eficiência, confiança e reprodutibilidade da investigação,impulsionar a colaboração interdisciplinar e fortalecer o envolvimento social e o impacto no mundo real.

1.2. Introdução à Ciência Aberta e Dados Abertos

A Ciência Aberta (CA) é um movimento que visa tornar o processo de investigação (métodos, dados, publicações, software) acessível a todos. A CA é uma abordagem baseada no trabalho cooperativo aberto e na partilha sistemática de conhecimento e ferramentas o mais cedo e amplamente possível no processo, e promove o acesso livre e transparente a todos os processos e resultados da investigação científica – desde os dados e publicações até os métodos e software – com o propósito de reforçar a colaboração, a reprodutibilidade e o impacto social da ciência.

O princípio orientador da Ciência Aberta é: "***Tão aberto quanto possível, tão fechado quanto necessário***". Isto significa que se deve partilhar dados de investigação logo que possível, sempre que seja legalmente e operacionalmente viável. O princípio ***"as open as possible, as closed as necessary"*** deve ser seguido, tornando os dados abertos por omissão. Em casos como dados relevantes para dual use (civil e militar), deve-se seguir as práticas de research security, garantindo que a abertura de dados científicos seja equilibrada com a proteção de informações sensíveis. Para além da eficiência académica, a CA promove a justiça social ao eliminar as barreiras financeiras para investigadores com menos recursos, ou em países em desenvolvimento, que de outra forma não teriam acesso a literatura científica de ponta. Economicamente, o acesso sem restrições permite que pequenas e médias empresas (PMEs) e start-ups acedam a resultados de investigação fundamental para inovar, sem os custos proibitivos de subscrições individuais, acelerando a transferência de tecnologia para o mercado.

O Movimento de Ciência Aberta baseia-se em valores como qualidade, integridade e reprodutibilidade, e representa a mudança de um modelo fechado para um modelo aberto, sendo impulsionado pela necessidade de maior transparência, rigor e eficiência na investigação. Baseia- se no trabalho cooperativo aberto e na partilha sistemática de conhecimento e ferramentas o mais cedo e amplamente possível, sendo um conceito inclusivo e relevante para todas as áreas e disciplinas.

Definições essenciais da CA

A CA inclui o Open Access (OA) a publicações, Dados FAIR, Software Open Source, Open Peer Review e Citizen Science. Isto inclui pre-registration e registered reports. A CA abrange o ciclo completo da investigação, desde o planeamento até à publicação.

Objetivos e Contexto Estratégico: O objetivo final da Ciência Aberta é aumentar a qualidade, a eficiência e a transparência da investigação. A correta implementação da RDM é crucial para a competitividade institucional, especialmente no contexto do programa Horizonte Europa (HE). A CA é uma prioridade política da Comissão Europeia, e a urgência da partilha de dados e resultados na pandemia de COVID-19 demonstrou a sua importância. A CA é baseada em valores como qualidade, integridade e reprodutibilidade.

Os cinco pilares chave da CA (UNESCO) são: Conhecimento científico aberto, infraestrutura científica aberta, comunicação científica, envolvimento aberto de atores sociais, e diálogo aberto com outros sistemas de conhecimento. De acordo com a Recomendação da UNESCO sobre Ciência Aberta, este movimento deve também focar-se na abertura à diversidade de conhecimentos, integrando saberes de comunidades marginalizadas e detentores de conhecimentos tradicionais. Além disso, a Ciência Aberta moderna exige a investigação e inovação responsáveis, garantindo que os avanços científicos sejam eticamente aceitáveis e socialmente desejáveis, promovendo o diálogo constante entre cientistas e decisores políticos.

Dados de Investigação (Research Data): Registos factuais recolhidos, observados ou gerados durante o processo de investigação, que servem de base para a análise e validação de resultados científicos. Exemplos incluem medições laboratoriais, questionários ou imagens. O conceito abrange software, algoritmos, protocolos, modelos, workflows e electronic notebooks.

Dados Abertos: São dados de investigação públicos, disponibilizados em formato acessível e reutilizável, sem restrições indevidas de acesso ou uso, mas sempre respeitando princípios éticos e legais (como privacidade e confidencialidade).

Dados em bruto (Raw Data): são dados originais recolhidos, mas ainda não elaborados ou analisados.

Conjunto de Dados (Dataset): Coleção estruturada de dados relacionados, geralmente organizada em formato tabular, numérico, textual, visual ou multimédia, e acompanhada

dos respetivos metadados.

Metadados (Metadata): São o rótulo dos dados, ou seja, informação descritiva sobre os dados que permite compreender, localizar e reutilizar um conjunto de dados. Exemplos de metadados são o título, autor, data de criação, metodologia e licença de uso. Metadados devem ser machine- actionable, padronizados e detalhados, incluindo proveniência e licenciamento.

Repositório de Dados (Data Repository): Plataforma digital segura onde os conjuntos de dados e metadados são depositados, preservados e tornados acessíveis ao público ou a grupos autorizados, devendo garantir a preservação a longo prazo e a interoperabilidade. Exemplos incluem Zenodo, Figshare e o Repositório ULisboa.

Ferramentas de Mapeamento Visual: Para além dos repositórios tradicionais, a descoberta de conhecimento pode ser potenciada por ferramentas de mapeamento visual, como o Open Knowledge Maps. Esta plataforma cria representações gráficas interativas de tópicos de investigação, permitindo que investigadores identifiquem rapidamente áreas disciplinares relacionadas e conceitos-chave, facilitando a navegação em domínios desconhecidos ou interdisciplinares sem barreiras.

Identificador Persistente (PID): Identificador único e duradouro, como o DOI (Digital Object Identifier) para o dataset e o ORCID (Open Researcher and Contributor ID) para o investigador. PIDs também se aplicam a instituições (e.g., ROR), software e subvenções.

Licenças Abertas: Instrumentos legais que definem como os dados e publicações podem ser usados, partilhados e adaptados, como a Creative Commons (CC BY, CC0).

Acesso Aberto (Open Access): Disponibilização gratuita e imediata de publicações científicas online, permitindo a leitura, descarregamento, cópia e partilha por qualquer pessoa, respeitando a autoria e integridade do trabalho. Enfatiza a abertura, colaboração e partilha precoce de conhecimento, resultados e ferramentas.

Open Innovation (Inovação Aberta): Abordagem de investigação que enfatiza a abertura, colaboração e o compartilhamento antecipado (early sharing) de conhecimento, resultados e ferramentas. Consiste no uso de fluxos de entrada e saída de conhecimento para acelerar a inovação interna e externa. Envolve stakeholders externos (indústria, cidadãos, decisores políticos) em processos de co-criação para acelerar o impacto da investigação.

Citizen Science (Ciência Cidadã): inclui o envolvimento de cientistas não profissionais na recolha de dados e na sua análise, é uma forma de envolvimento ativa e recomendada. A Citizen Science implica ter projetos que envolvem ativamente o público em qualquer fase da investigação, atuando como colaboradores, contribuidores ou líderes de projeto. O tipo de atividades inclui p.ex. Codesign: Stakeholders ajudam a desenhar o processo de investigação (Ex: Pacientes co-desenham um protocolo de estudo clínico); Co-creation: Abordagem colaborativa onde stakeholders ativamente moldam o projeto, do design à implementação (Ex: Agricultores co-desenvolvem soluções de agricultura sustentável); e Co-assessment: Stakeholders avaliam os resultados para garantir relevância e usabilidade (Ex: Policymakers envolvidos na revisão de modelos de impacto climático).

Os projetos de Ciência Cidadã devem ser desenhados considerando a Escala de Haklay sobre os níveis de participação: 1. Crowdsourcing (cidadãos como sensores); 2. Inteligência Distribuída (cidadãos como intérpretes básicos de dados); 3. Ciência Participativa (envolvimento na definição do problema); e 4. Ciência Cidadã Extrema (colaboração total na análise e tomada de decisão). Definir o nível de envolvimento desde o início ajuda a gerir expectativas e a desenhar estratégias de motivação adequadas para os voluntários.

Preprints: Uma estratégia fundamental para acelerar a disseminação é o uso de Preprints. O depósito de versões manuscritas antes da revisão por pares em servidores como arXiv ou bioRxiv permite o feedback imediato da comunidade. Além disso, os autores devem estar cientes do "Direito de Retenção de Direitos" (Rights Retention Strategy), que permite aos investigadores manterem direitos suficientes sobre os seus manuscritos para os depositarem em repositórios institucionais sem períodos de embargo, independentemente da política da editora.

A estrutura principal de qualidade na Ciência Aberta são os princípios FAIR (Findable, Accessible, Interoperable, Reusable), vão além da CA concentrando-se na qualidade técnica dos dados:

Findable (F): requer Identificadores Persistentes (PIDs, como DOI e ORCID) e metadados ricos, que sejam pesquisáveis e localizáveis online;

Accessible (A): significa que os dados são acessíveis sob condições claras, podendo ser restritos (managed access) se necessário, e podem ser restritos e continuar FAIR. A acessibilidade deve ser garantida através de um protocolo que seja aberto, gratuito e universalmente implementável, e que permita procedimentos de autenticação e

autorização;

Interoperable (I): exige formatos abertos e standards de vocabulário (ex: CSV em vez de XLSX). Os dados devem ser disponibilizados em formatos sem perda (lossless), e deve-se usar definições partilhadas e termos padronizados dentro do domínio específico;

Reusable (R): foca-se na documentação detalhada (proveniência) e licenças claras (p.ex. CC BY, CC0). A documentação deve ser bem elaborada (e.g., README files), incluindo proveniência e ferramentas/instrumentos necessários para reproduzir os resultados.

O FAIR é a base para iniciativas como a European Open Science Cloud (EOSC). Os quatro princípios, em maior detalhe, são os seguintes:

1. F (Findable) – Localizável: Requer o uso de Identificadores Persistentes (PIDs) como DOI (para o dataset) e ORCID (para o autor). Os metadados devem ser ricos para permitir a descoberta. PIDs garantem que os dados podem ser localizados e citados de forma estável, e Repositórios confiáveis (Trusted Repositories) devem atribuir consistentemente PIDs. PIDs também se aplicam a instituições (e.g., ROR), subvenções, software e hardware.

2. A (Accessible) – Acessível: Não significa necessariamente aberto, mas acessível sob condições claras. O princípio orientador é "Tão aberto quanto possível, tão fechado quanto necessário". A acessibilidade é garantida ao saber onde os dados estão e como lhes aceder (protocolo). Se os dados forem restritos (p.ex., por RGPD ou segredos comerciais), deve existir um procedimento de acesso gerido (managed access procedure), se necessário com autenticação e autorização.

3. I (Interoperable) – Interoperável: Exige o uso de vocabulários, standards e formatos abertos para permitir a integração e o processamento por máquinas. Os dados devem ser disponibilizados em formatos não-proprietários (Ex: CSV ou Parquet em vez de XLS). Os dados devem ser disponibilizados em formatos sem perda (lossless). Metadados e dados devem ser descritos usando termos normalizados. Ferramentas como o FAIRsharing ajudam a encontrar os padrões de metadados específicos de cada domínio.

4. R (Reusable) – Reutilizável: Requer o uso de licenças claras e atribuição adequada, preferencialmente CC BY ou CC0. Os dados devem ser acompanhados de informações sobre condições de uso, proveniência detalhada e documentação. A proveniência deve ser documentada com um audit trail permanente. A proveniência (histórico de geração e processamento) deve ser documentada nos metadados. Para uma reutilização

sustentável, os dados devem ser convertidos para formatos de arquivo como PDF/A para evitar a sua obsolescência.

Princípios CARE: uma estrutura ética essencial para dados científicos e conhecimento de comunidades locais e indígenas, onde o Collective Benefit e a Authority to Control são centrais. São recomendados para dados comunitários ou sensíveis:

- Collective Benefit: Os dados devem beneficiar as comunidades locais e indígenas, e não apenas objetivos académicos.
- Authority to Control: Definição explícita de quem tem o direito de controlar o acesso e as cláusulas de Propriedade Intelectual (PI), em particular comunidades locais e indígenas.
- Responsibility: Responsabilidade de transparência no uso de dados, e perante quem fornece os dados.
- Ethics: Gestão de dados de forma ética.

O Papel da GDI (Research Data Management, RDM): A GDI é a prática de apoio que assegura a qualidade e a reutilização dos dados, abrangendo o ciclo de vida completo, desde o planeamento até à preservação. A RDM correta aumenta a eficiência da pesquisa e o valor dos dados, evitando duplicações. O documento PDG (Plano de Gestão de Dados) descreve como os dados de investigação serão geridos durante e após o projeto: criação, armazenamento, partilha, preservação e acesso futuro.

1.3. Estratégia Institucional e Contexto Legal

Estratégia da ULisboa: A ULisboa promove ativamente os princípios da CA, valorizando o FAIR e a inovação aberta. A ULisboa é membro da Unite! Alliance, que dá muita ênfase à CA. O desenvolvimento do Centro iRe:SEARCH visa elaborar a Política Institucional de Dados, formações, documentos de suporte à comunidade ULisboa, e um canal de suporte técnico. A ULisboa aderiu ao Acordo CoARA em 2024. Elementos-chave a adotar incluem a certificação de repositórios, a garantia da preservação de dados a longo prazo e o papel dos Data Stewards (Gestores de Dados). Os repositórios devem ser transparentes sobre a governação, sustentabilidade financeira e plano de continuidade.

No âmbito da reforma da avaliação da investigação CoARA, está a ganhar força o uso de Currículos Narrativos. Em vez de uma lista exaustiva de publicações e fatores de impacto, o investigador descreve qualitativamente o seu contributo para a geração de conhecimento, para a formação de equipas e para a sociedade, permitindo que atividades de Ciência Aberta (como a curadoria de um excelente dataset) tenham o mesmo peso que um artigo numa revista de prestígio.

Compromisso Nacional: O compromisso político português (Resolução do Conselho de Ministros n.º 21/2016 e Lei da Ciência, DL n.º 63/2019) reforça a CA. Portugal tem um compromisso político com a CA visando que as instituições de I&D ativamente contribuam para a CA, assegurando acesso livre ao conhecimento científico. A Política de Acesso Aberto da FCT (em vigor desde fev/2025) obriga ao cumprimento de requisitos FAIR, incluindo a preservação de dados por pelo menos 10 anos e a disponibilização de metadados em CC0 (Domínio Público). Além disso, a política da FCT não admite períodos de embargo em conteúdos de acesso aberto e exige o uso da estratégia de retenção de direitos para publicações. A FCT tem sido fundamental para alinhar as políticas nacionais com a política europeia de CA, e a retenção de direitos é um mecanismo chave para o depósito imediato em acesso aberto do Manuscrito Aceite do Autor (AAM).

Relevância do RGPD: O RGPD (Lei n.º 58/2019) é o principal condicionante legal. O RGPD é aplicado a dados recolhidos do domínio público, onde a pessoa pode ser identificada, e a dados pseudonimizados. Garante que o princípio da Acessibilidade (A de FAIR) não viole a Privacidade. Exige a implementação da proteção desde a conceção e o Data Protection by Design, e impõe princípios como a minimização de dados e a transparência. Isto inclui a Minimização de Dados, Limitação da Finalidade, Integridade e Confidencialidade e Accountability. É obrigatória uma Avaliação de Impacto sobre a Proteção de Dados (DPIA) para tratamento de dados sensíveis em larga escala. Exemplos de alto risco incluem a monitorização sistemática de áreas públicas em larga escala, utilização de novas tecnologias (IA/ML), e o tratamento em larga escala de categorias especiais de dados.

Gestão de Dados Sensíveis: Dados sensíveis devem ser geridos com salvaguardas, como anonimização/pseudonimização e acesso controlado, mesmo que os metadados sejam abertos. A anonimização é o processo de remoção de informações de identificação pessoal, sendo irreversível, o que retira os dados do âmbito do RGPD, permitindo a partilha aberta. A Anonimização retira os dados do âmbito do RGPD, enquanto a Pseudonimização é reversível, mantendo os dados no âmbito do RGPD. O Acesso controlado é uma salvaguarda crucial que permite a partilha e reutilização de dados de investigação sob condições claras, mesmo quando esses dados não podem ser totalmente

abertos. A ferramenta AMNESIA do OpenAIRE é um recurso open source que pode ser usado livremente para a anonimização e pseudonimização de dados sensíveis.

1.4. Caso de Estudo: Implementação dos Princípios FAIR

O Caso de Estudo consiste na análise de um projeto de Smart City Data, com dados de sensores (qualidade do ar e trânsito), uma publicação sobre padrões de fluxo e um software aberto para visualização 3D dos dados.

Aceleração da Descoberta: Ligar o FAIR à aceleração da descoberta é crucial, especialmente com grandes volumes de dados (Big Data). A reutilização dos dados, facilitada pelo FAIR, permite evitar duplicações e acelera o progresso da ciência.

Exemplos F (Findable) e A (Accessible):

- Findable: Depósito no Zenodo (trusted repository), que atribui DOI e exige metadados básicos (título, autor, data, licença). Uso de ORCID para autores. Os PIDs garantem que os dados e os autores sejam referenciados de forma estável.
- Accessible: Os dados e metadados devem ser acessíveis através de protocolos standard, abertos e seguros, respeitando políticas de acesso e privacidade.

Exemplos I (Interoperable) e R (Reusable):

- Interoperable: Uso de vocabulários controlados e formatos abertos (Ex: CSV para os dados dos sensores). Uso de standards de metadados como os elementos Dublin Core (Identifier, Subject, Type, Format, and Coverage). Também podem ser usadas ontologias CityGML e INSPIRE para compatibilidade com dados urbanos europeus.
- Reusable: Aplicação de uma licença clara (p.ex., CC BY para dados, CC0 para metadados e MIT License para o software 3D). Para reutilização completa, deve ser partilhado o código e o software necessários (Ex: software de visualização no GitHub). Deve ser fornecido um Codebook (dicionário de dados) para contextualizar o dataset. Documentação completa com ficheiros README.md com descrição técnica e metodológica, versões controladas, e dados anonimizados.

Infraestrutura Institucional: O ULisboa Research Portal / SIIC ULisboa é a plataforma central onde os dados devem ser carregados inicialmente, assegurando a interoperabilidade com RCAAP e OpenAIRE. Qualquer infraestrutura de repositórios deve ter ou visar a certificação de confiabilidade (Trustworthy Data Repositories, ex: CoreTrustSeal), incluindo demonstrar transparência na governação, sustentabilidade financeira e plano de continuidade.

Resumo FAIR neste Caso de Estudo:

- Findable: Datasets com DOI e metadados normalizados (SensorML, Dublin Core).
- Accessible: Acesso via HTTPS e API aberta; licenças claras; metadados sempre disponíveis.
- Interoperable: Uso de formatos abertos (CSV, GeoJSON), ontologias CityGML e INSPIRE.
- Reusable: Documentação completa (ficheiros README.md), versões controladas, dados anonimizados e licenças abertas.

1.5. Sistemas e Repositórios de Ciência Aberta

Os sistemas e repositórios de Ciência Aberta podem ser classificados em seis tipos:

1. Repositórios Institucionais (Escala Local): Mantidos por universidades ou centros de investigação, servem para armazenar e disponibilizar os resultados científicos produzidos internamente. Interoperam com plataformas como OpenAIRE, fornecendo dados à rede. Exemplos: Apollo, Aaltodoc, Repositório ULisboa.

2. Repositórios Temáticos (Escala Europeia/Global): Servem comunidades científicas específicas, armazenando dados e publicações num domínio concreto, permitindo a especialização disciplinar. Permitem interoperabilidade internacional via DOI e metadata harvesting (OAI-PMH), ligando-se à EOSC. Exemplos: Zenodo, PANGAEA, ENA.

3. Repositórios Nacionais (Escala Nacional): Coordenados a nível de país, agregam os repositórios institucionais e temáticos, atuando como nós nacionais da European Open Science Cloud (EOSC). Exemplos: HAL, RCAAP, NORA.

4. Repositórios Internacionais e Plataformas Pan-Europeias (Escala Europeia): Conectam instituições e países sob uma infraestrutura comum, promovendo a federação de serviços e a interoperabilidade técnica e semântica. Promovem interoperabilidade técnica (via APIs, DOI, ORCID) e semântica (via esquemas de metadados como Dublin Core, DataCite, CERIF). Exemplos: EOSC, OpenAIRE. O OpenAIRE é uma rede descentralizada e um ecossistema que consolida toda a informação num único Grafo Semântico (OpenAIRE Graph), ligando metadados de mais de 70.000 fontes académicas.

5. Repositórios de Software e Código Aberto (Escala Global): Parte essencial da Ciência Aberta, garantindo transparência nos métodos e reprodutibilidade científica. Contêm software e scripts usados em investigação, com controlo de versões. Exemplos: GitHub, Software Heritage. O Software Heritage visa arquivar e preservar o código-fonte de todo o software publicamente disponível, garantindo a preservação a longo prazo.

6. Sistemas de Apoio e Interoperabilidade (Escala Institucional/Global): Complementam os repositórios, integrando informação sobre projetos, publicações, dados e investigadores (ex. PURE), infraestruturas para PIDs e agregadores de metadados de vários repositórios (ex. OpenAIRE Graph). Exemplos: PTCRIS, ORCID, DOI, Re3data. O Global Re3data é um registo que ajuda investigadores a encontrar repositórios de dados confiáveis, permitindo a pesquisa por repositórios registados no OpenAIRE ou que são certificados.

1.6. Obrigações e Boas Práticas

Obrigatório:

- Seguir princípios FAIR e garantir preservação por 10 anos.
- Usar identificadores persistentes (DOI, ORCID) e formatos abertos.
- Publicar só em trusted repositories. Repositórios não são considerados trusted se forem websites ou serviços cloud como o Dropbox ou ResearchGate.
- Sem períodos de embargo para publicações em acesso aberto. O acesso imediato a publicações peer-reviewed é exigido no Horizonte Europa, e para tal, os autores devem reter direitos suficientes de PI.

- Usar licenças claras e abertas. Os metadados devem ser abertos e licenciados sob CC0.

- Proteção e Ética dos Dados: seguir RGPD e princípios éticos no tratamento de dados pessoais (p.ex., anonimização e consentimento informado). A adesão aos princípios éticos é tão vinculativa quanto a lei no contexto da investigação científica.

Boas práticas adicionais:

- Usar mais do que um repositório, preservação por mais de 10 anos.

- Seguir também princípios CARE, especialmente para dados comunitários e sensíveis.

- Associar dados, publicações, e software: datasets, artigos e código ligados com PIDs (DOI, ORCID, ROR) e descrições cruzadas nos metadados. Ter uma Data Availability Statement (DAS) em todos os artigos, garantindo que o dataset é Findable e Accessible. A DAS deve informar a localização e as condições de acesso e não deve instruir os leitores a contactar o autor para obter os dados.

- Documentar com rigor (p.ex., ficheiros README detalhados, versões do dataset, metodologias e scripts de tratamento).

- Monitorizar reutilização e impacto (p.ex., downloads, citações e reutilização dos dados –

  altmetrics).

2. Gestão e Preservação de Dados Abertos, Condicionalismos Legais e Regulamentares

2.1. Conceitos Chave

Este guia centra-se na preservação a longo prazo dos dados, que é essencial para a reprodutibilidade e exige a retenção mínima de 10 anos (requisito da FCT). Este requisito decorre da obrigação de armazenar dados pelo menos 10 anos após a publicação. A preservação sustentável requer a conversão dos dados para formatos de arquivo abertos (p.ex. ficheiros CSV, texto simples, PDF/A, TIFF, ODT) para evitar a obsolescência tecnológica e garantir a interoperabilidade. É essencial usar formatos abertos e lossless (sem perda de qualidade) que retenham todos os dados e sejam acessíveis em várias plataformas.

Durante um projeto de investigação, o armazenamento seguro deve ser feito em servidores institucionais ou nuvem aprovada, sendo desencorajado o uso de discos locais e de pen drives. O depósito deve ser feito em Repositórios Confiáveis que cumpram padrões internacionais como a certificação CoreTrustSeal. Os Repositórios Confiáveis devem assegurar a precisão, integridade, autenticidade e acesso dos conteúdos e usar metadados detalhados, normalizados e machine- actionable (incluindo proveniência e licenciamento).

Os planos de backup e planos de recuperação são cruciais, e a Regra 3-2-1 é o padrão de ouro. A eliminação de dados (especialmente pessoais e intermédios) após o período de retenção (10 anos) deve seguir a legislação aplicável (RGPD).

As exceções para manter os dados fechados ("*as closed as necessary*") devem ser justificadas de forma clara, abrangendo: valor comercial (segredos comerciais, proteção de PI); preocupações de segurança (por exemplo, projetos relacionados com a defesa); interesses estratégicos da UE (proteção da autonomia/segurança).

Cada financiador pode ter ainda obrigações e condicionalismos legais adicionais, por exemplo no Horizonte Europa, há obrigações relativas ao Data Management Plan (DMP): é obrigatório criar e submeter o DMP, sendo que a primeira versão é um deliverable formal a ser apresentado até ao Mês 6 do projeto. A primeira versão curta (menos de meia página) do DMP já é necessária na fase de proposta. No caso de emergências públicas, a autoridade financiadora pode solicitar o acesso imediato aos outputs de

investigação, o qual deve ser fornecido sob licença CC BY ou CC0. No caso de emergências públicas (e se solicitado p.ex. pela UE), os outputs de investigação devem ser depositados num repositório imediatamente, com licenças não exclusivas, sob termos justos, por até quatro anos após o fim da ação, se a abertura total não for possível.

2.2. Melhores Práticas em Gestão de Dados a Longo Prazo

A Preservação Sustentável é fundamental para o princípio da Reutilização (R de FAIR). A preservação sustentável exige a retenção dos dados por, pelo menos, 10 anos, ou que se justifique um período diferente. Para evitar a obsolescência tecnológica, é crucial converter os dados para formatos de arquivo abertos (como PDF/A, TIFF ou CSV simples), sendo os formatos abertos, como CSV, preferíveis a formatos proprietários, como XLSX. O uso de formatos proprietários só é aceitável se houver consenso generalizado na comunidade, mas mesmo nestes casos deve ser produzida uma cópia em formato aberto.

A preservação começa com a Documentação e Metadados. O conjunto de dados (dataset) deve ser compreensível no futuro, o que exige metadados completos sobre o contexto, métodos e estrutura. Estes metadados devem ser descritivos, machine-actionable e padronizados, e devem incluir informação detalhada sobre a proveniência, que é a origem e história dos dados. A FCT exige que o Plano de Gestão de Dados (PGD) identifique o(s) standard(s) de metadados e o(s) vocabulário(s) a serem utilizados. A documentação deve ser criada e mantida em todas as fases do ciclo de vida dos dados, cobrindo métodos, protocolos, ficheiros de dados e resultados preliminares. Deve também descrever o contexto em que os dados foram criados.

O Controlo de Qualidade, Integridade e Proveniência (RDM) é essencial para garantir a confiança nos dados e demonstrar a ausência de fraude. O Controlo de Qualidade envolve processos como calibração, amostras ou medições repetidas, padronização da captura de dados, validação de entrada de dados e peer review dos dados. A Proveniência deve ser registada sistematicamente, incluindo metadados detalhados sobre as transformações, versões e o registo histórico dos dados (tempo, operações e parâmetros). O Controlo de Qualidade deve seguir um procedimento padrão da organização.

A seleção de Repositórios Confiáveis é crucial para a preservação a longo prazo. Estes Trusted Repositories devem cumprir padrões internacionais, como a certificação CoreTrustSeal, e ser estáveis. Eles devem ser transparentes sobre a governança, sustentabilidade financeira, período de retenção e plano de continuidade, e devem garantir a integridade, autenticidade e acesso dos conteúdos. Os repositórios confiáveis devem gerir a preservação dos dados de forma documentada, detalhando a sua missão, âmbito, e plano de contingência A Science Europe estabelece critérios para repositórios, incluindo a atribuição de Identificadores Persistentes (PIDs).

As políticas de RDM, como o mandato da FCT, exigem a preservação de dados por, pelo menos, 10 anos. Os repositórios confiáveis devem atribuir um PID único (como o DOI) a cada dataset para garantir a identificação e citação estável. Os PIDs são essenciais para que os dados sejam Findable (F), e o repositório deve suportar o versionamento dos dados e o rastreamento da sua proveniência.

2.3. Estratégias de Armazenamento e Preservação

A preservação digital garante que os dados permaneçam acessíveis e utilizáveis a longo prazo, protegendo contra a obsolescência de formatos e a degradação de suportes.

Durante o projeto, o armazenamento deve ser seguro e em rede (servidores institucionais ou nuvem aprovada), sendo desencorajado o armazenamento em discos rígidos locais ou pen drives. Se forem utilizados dispositivos portáteis, deve ser garantido que existem cópias em drives de rede e backup de armazenamento. O armazenamento deve ser feito em espaços internos ou aprovados pela instituição. O PGD deve descrever onde os dados serão armazenados e os seus backups (preferencialmente automáticos). Para projetos colaborativos, é fundamental utilizar diretórios de grupo partilhados com amplo acesso para garantir que os dados permaneçam com a organização e não com o indivíduo.

Para planos de backup e recuperação, deve ser adotada a Regra 3-2-1 de Backup como padrão de ouro: 3 cópias, 2 tipos de média e 1 cópia off-site. A política de backup deve ser clara, incluindo o tipo e a frequência, e um plano de recuperação para os dados em caso de incidente. Por exemplo, os Aalto Network Drives usam um sistema de backup automático baseado em snapshots horários. Os serviços de cloud (OneDrive, Teams, Google Drive) usam um esquema de replicação onde os dados são armazenados em diferentes data centers e sistemas de armazenamento.

A Integridade e a Reprodutibilidade dependem do registo sistemático da origem, o que exige metadados detalhados sobre as transformações, versões e o registo histórico (tempo, operações e parâmetros). O repositório deve suportar o versionamento dos dados para rastreabilidade e controlo de qualidade.

Para mitigar riscos técnicos, recomenda-se a implementação de verificações de integridade (comom *Fixity Checks* ou *Checksums*) periódicas, que detetam automaticamente a corrupção de bits (*bit rot*).

As políticas de RDM devem incluir diretrizes claras para a Destruição Segura dos Dados. Após o período de retenção exigido, a eliminação deve cumprir a legislação aplicável, como o RGPD. A eliminação de dados intermédios (que não se destinam a publicação) deve ser planeada no PGD, em alinhamento com a declaração de privacidade e com o princípio da minimização de dados.

2.4. Conformidade Legal e Regulamentar

O RGPD e a gestão de Dados Sensíveis são cruciais. A acessibilidade dos dados (A de FAIR) não pode violar a Privacidade. Os dados confidenciais exigem proteção e segurança reforçada, devendo ser encriptados e o acesso deve ser controlado e registado (logging). As causas para o fecho ou restrição de dados identificadas no PGD incluem PI, confidencialidade e segurança nacional.

Em termos de Segurança da Investigação, PI e NDAs, os dados confidenciais são limitados a pessoas específicas. Em projetos colaborativos, a PI e o controlo de acesso devem ser cobertos pelo Acordo de Consórcio. A partilha de dados e/ou exceções devem estar em contratos e Acordos de Não Divulgação (NDAs/MoUs) com entidades privadas. Para dados que requerem ambientes de armazenamento muito estritos (como os de Dual Use – potencial uso militar da tecnologia), o PGD deve especificar ambientes de armazenamento seguros (encriptados e com acesso restrito).

O RGPD exige a implementação da Proteção de Dados Desde a Conceção e por Padrão (Data Protection by Design and Default). As restrições legais e éticas devem ser consideradas desde o início do projeto. Projetos de alto risco ou que tratem dados sensíveis em larga escala exigem uma Avaliação de Impacto sobre a Proteção de Dados (AIPD ou DPIA). O RGPD exige a accountability (manter registos). A legislação aplicável no

ordenamento jurídico português inclui as Leis n.º 58/2019 e n.º 59/2019, a Lei n.º 34/2009, a Portaria n.º 312-A/2022 e a Resolução do Conselho de Ministros n.º 41/2018. A Avaliação de Impacto (DPIA) é exigida em casos de utilização de novas tecnologias (IA, machine learning), monitorização sistemática de áreas públicas em larga escala, e tratamento de dados de menores ou dados sensíveis.

A AIPD é exigida quando o tratamento é de alto risco, por exemplo, em casos de monitorização sistemática de áreas públicas em larga escala, utilização de novas tecnologias (IA, machine learning), tratamento de dados de menores ou dados sensíveis, criação de perfis (profiling) ou decisões automatizadas com impacto legal, e tratamento em larga escala de categorias especiais de dados (como raciais, étnicos ou genéticos). As medidas de mitigação incluem salvaguardas técnicas e organizacionais para reduzir riscos. O PGD deve detalhar o cumprimento do RGPD para dados pessoais, incluindo a forma como será gerida a propriedade dos dados e os direitos de PI.

Os Acordos de Partilha de Dados (DSAs) são cruciais para dados que exigem acesso controlado e devem prever o registo (logging) de acesso.

As Etapas da AIPD incluem:

- Descrever o tratamento (Natureza, âmbito, finalidade e base legal);
- Avaliar a necessidade e proporcionalidade;
- Identificar riscos (acesso não autorizado, perda ou uso indevido de dados);
- Definir medidas de mitigação;
- Demonstrar conformidade, documentando o processo como prova de cumprimento do RGPD.

O PGD é o documento ideal para gerir questões de PI. Deve ser definido quem será o proprietário dos dados (quem terá os direitos de controlar o acesso).

O PGD é obrigatório (até ao mês 6 no Horizon Europe) e deve ser dinâmico, com atualizações regulares. O PGD da FCT exige que se detalhe como será assegurado o cumprimento da legislação sobre dados pessoais (e processos de anonimização ou pseudonimização, se aplicáveis). Deve cobrir explicitamente o plano de backup, a segurança de dados sensíveis e a licença de partilha.

2.5. Caso de Estudo: Desafios de Preservação (Smart City)

No exemplo de um projeto Smart City (dados de sensores de qualidade do ar, trânsito e ruído), a Anonimização Robusta é crucial, pois retira os dados do âmbito do RGPD, permitindo a partilha aberta (R de FAIR). A Pseudonimização (p.ex., encriptação) é reversível e mantém os dados sob o RGPD. A ferramenta AMNESIA do OpenAIRE pode ser utilizada para anonimizar dados. As funcionalidades do AMNESIA incluem a generalização de valores (substituindo o valor específico por um agregado através de hierarquias) e a supressão de registos. O AMNESIA pode ser utilizado localmente para maior segurança, evitando a transferência de dados pessoais através da Internet.

Para dados que não podem ser abertos, o Acesso Controlado é a solução (A de FAIR), sendo que o acesso é restrito através de um procedimento de acesso gerido (managed access procedure). É necessária a formalização de Acordos de Partilha de Dados (DSAs) para regular o acesso controlado a colaboradores. No caso de uso de câmaras de vídeo que capturam dados de trânsito, é necessário realizar um DSA com as entidades que geram os dados e realizar uma DPIA para dados de vídeo.

Os Custos de RDM (curadoria, tempo de RHs, armazenamento, encargos de repositório) podem ser elegíveis para financiamento (ex: no HE). O PGD deve estimar recursos (financeiros e de tempo), incluindo os custos de preservação a longo prazo (10 anos após o final do projeto) e os decorrentes de seguir a regra 3-2-1 de Backup. Os custos elegíveis para financiamento no Horizonte Europa incluem custos de armazenamento, hardware, tempo de pessoal, custos de preparação dos dados para depósito e encargos de repositório.

2.6. Obrigações e Boas Práticas

Obrigatório:

- É obrigatório atualizar o PGD em caso de mudanças significativas. O DMP deve ser atualizado regularmente para refletir as decisões que são implementadas.
- Os dados devem ser depositados num Repositório Confiável.

- É mandatório preservar os dados por um período mínimo (Ex: 10 anos, segundo a FCT).
- O RGPD deve ser integralmente cumprido no tratamento de dados pessoais e sensíveis (o acesso a estes dados pode ser restrito).
- Garantir o Backup dos dados e um plano de recuperação. O PGD deve descrever a política e a frequência de backup.
- O PGD deve detalhar a alocação de recursos e custos para a gestão e preservação dos dados.
- Os Metadados devem ser FAIR, acessíveis e licenciados sob CC0 (Domínio Público). Deve-se disponibilizar softwares/ferramentas se forem necessários para a reutilização dos dados.

Boas práticas adicionais:

- Recomenda-se utilizar PGD Machine-Actionable (maDMP) para facilitar a interoperabilidade com repositórios. É recomendado que o DMP seja público.
- É uma boa prática depositar em repositórios que possuam Certificação CoreTrustSeal para garantir a preservação a longo prazo.
- Outras boas práticas incluem fornecer documentação detalhada, p.ex. README files, codebooks (livro de código ou dicionário de dados), para garantir a reutilização, formalizar o Acesso Controlado através de Acordos de Partilha de Dados (DSA) e implementar o Backup dos dados (regra 3-2-1) com armazenamento institucional.
- A ferramenta AMNESIA (ou similar) deve ser utilizada para anonimização robusta.

3. Direitos de Autor e Direitos Conexos

3.1. Conceitos Chave

Este guia distingue entre Direitos de Autor (DA), que protegem a expressão criativa, como texto, publicações e código-fonte, e Direitos Sui Generis, que protegem o investimento na criação e curadoria de conjuntos de dados (datasets) brutos, cuja propriedade é geralmente da instituição. Os Direitos Conexos protegem o investimento substancial feito na organização dos conteúdos de bases de dados ou compilações. Os outputs de investigação, que são os resultados aos quais pode ser dado acesso online, incluem publicações científicas, dados digitais, software, algoritmos, protocolos, modelos, workflows e electronic notebooks. O PGD é o local ideal para definir a Propriedade Intelectual (PI), a titularidade e os direitos de controlo de acesso. São ainda abordadas as opções de licenciamento (e.g., Creative Commons, CC) que promovem a partilha e reutilização de resultados.

Para publicações resultantes de projetos financiadas pela FCT ou Horizon Europe, é obrigatório garantir o Acesso Aberto imediato (sem embargo) aos artigos científicos. A Estratégia de Retenção de Direitos (Retention Rights Strategy) é o mecanismo que autores devem usar para reter direitos sobre o Manuscrito Aceite do Autor (AAM), o que permite o depósito imediato em acesso aberto. Apesar de não ser admitido embargo, nestes projetos, preservar o segredo comercial é uma justificação válida para restringir o acesso, e a divulgação deve ser adiada até uma salvaguarda ser assegurada p.ex. até uma patente ser formalizada.

3.2. Fundamentos Legais e Aplicação

O Direito de Autor (DA) e os Direitos Conexos protegem diferentes aspetos da produção de investigação. O DA protege a forma de expressão de uma obra, como o texto de um artigo científico ou o código-fonte de um software, exigindo que a obra resulte de escolhas livres e criativas. Os Direitos Conexos, também conhecidos como sui generis database rights (uma proteção legal criada pela Diretiva de Bases de Dados), protegem o investimento substancial feito na organização dos conteúdos de bases de dados ou compilações, mesmo que o conteúdo em si não seja criativo.

Publicações e Software: Tipicamente protegidos por Direitos de Autor, sendo esta proteção geralmente da responsabilidade do autor.

Raw Data e Datasets: Dados Brutos (Raw Data) factuais por si só não são protegidos por Direitos de Autor, devido à falta de criatividade. Contudo, a estrutura e curadoria de um dataset ou base de dados podem ser protegidas pelo Direito sui generis. Esta proteção sui generis é, em regra, do investidor (empregador ou instituição), e não do investigador (autor). A reprodução de obras para Text and Data Mining (TDM) para fins de investigação científica não pode ser derrogada por contrato nem prevenida por meios técnicos, desde que a organização de investigação que realizao TDM tenha acesso legal à obra.

3.3. Outras Medidas de Proteção e Justificações para Restrição

Além do DA e dos Direitos Conexos, a Propriedade Intelectual (PI) pode ser protegida por Patentes para invenções com novidade absoluta. Nestes casos a publicação de resultados deve ser adiada até que o depósito da candidatura à patente seja formalizado. Outras medidas incluem Marcas (protegem a identidade de projetos) e Desenho Técnico (petty patent ou utility model).

O Segredo Comercial é uma justificativa válida para restringir o acesso, se a abertura comprometer a exploração ou a proteção da PI.

O Plano de Gestão de Dados (PGD) é crucial para gerir a PI. Deve detalhar quem é o proprietário dos dados e quem terá os direitos de controlar o acesso. Em Projetos Colaborativos, a PI (incluindo o direito sui generis) e o controlo de acesso devem ser cobertos pelo Acordo de Consórcio. É fundamental que o PGD indique se há restrições na reutilização de dados de terceiros, como aqueles cobertos por acordos de não divulgação (NDAs/MoUs).

As justificações válidas para fechar dados (restrição ou exceção) são:

- Quando a abertura possa impedir a proteção por patentes ou dificultar a exploração comercial (Segredos Comerciais), ou seja, no caso de dados comercialmente valiosos se a abertura comprometer a sua exploração ou

dificultar a proteção de PI;

- Obrigações legais/éticas, como o RGPD (dados pessoais/sensíveis);
- Segurança pública ou interesses da União Europeia/Segurança Nacional. A justificação do embargo, indicando o período de adiamento, deve estar no PGD.

3.4. Declaração de Disponibilidade de Dados (DAS)

A Declaração de Disponibilidade de Dados (DAS) é obrigatória em todos os artigos. Funciona como um elemento de transparência que garante os princípios Findable e Accessible do dataset.

A DAS deve informar explicitamente os leitores sobre a localização dos dados e as condições de acesso e reutilização. Não deve indicar que os leitores devem contactar o autor para obter os dados (o que seria um problema de acessibilidade, Accessibility issue). Esta falha na Acessibilidade não é permitida, pois um requisito é que os dados sejam acessíveis (A de FAIR). A DAS deve detalhar as condições e a localização (PID) do dataset, mesmo que este seja restrito e que apenas os Metadados estejam disponíveis. A COPE (Committee on Publication Ethics) fornece orientações sobre como escrever a DAS de forma ética e transparente. É possível mencionar o PGD na DAS, caso este esteja publicado.

3.5. Modelos de Acesso Aberto (AA) para Publicações

Existem vários modelos de Acesso Aberto para publicações científicas:

- Híbrida: Modelo misto (Subscrição + Acesso Aberto pago via APC). Mistura conteúdo aberto e fechado (paywall). O pagamento de APC é geralmente exigido para o AA. APCs para revistas híbridas (hybrid journals) não são elegíveis para reembolso no Horizonte Europa (HE).
- Gold Open Access (Completo): Acesso Aberto Total, com todo o conteúdo imediatamente aberto. Geralmente exige APC.

- Diamond Open Access: Acesso Aberto Total, com todo o conteúdo imediatamente aberto, mas nunca exige APC (é gratuito para autor e leitor). A Open Research Europe (ORE) é uma plataforma de publicação Diamond Open Access gratuita e financiada pela Comissão Europeia.

- Green Open Access (Via Verde): Auto arquivamento. O Manuscrito Aceite do Autor (AAM), revisto por pares, é depositado num repositório aberto, sem APCs. O Self- archiving é o processo de depositar uma versão do trabalho num repositório confiável.

APC é o Author Processing Charge (Encargo de Processamento de Autor) – A APC é um pagamento geralmente exigido para garantir o Acesso Aberto (AA) de artigos em alguns modelos de publicação.

Tipo de Revista	Modelo	APC para Autor	Acesso ao Conteúdo
Híbrida	Misto (Subscrição + AA Pago)	Sempre exigida para o AA	Uma mistura de conteúdo aberto e conteúdo fechado (paywall).
Gold Open Access (Completo)	Acesso Aberto Total	Geralmente exigida (apenas por ~29% das revistas Gold).	Todo o conteúdo é imediatamente aberto.
Diamond Open Access	Acesso Aberto Total	**Nunca exigida** (Gratuito para autor e leitor).	Todo o conteúdo é imediatamente aberto.
Green Open Access (Via Verde)	Auto arquivamento	Nenhuma.	O artigo é publicado numa revista, mas o manuscrito revisto por pares (AAM) é depositado num repositório aberto.

3.6. Opções de Licenciamento Aberto

O Licenciamento Aberto é necessário porque, sem uma licença, aplica-se a regra por omissão de "Todos os Direitos Reservados" ("*default rule all rights reserved*"), impedindo a reutilização. Uma licença é um contrato que concede permissões claras para usar um determinado trabalho (dados, código, texto). O licenciamento deve promover a reutilização, adaptação e redistribuição.

Licenças Creative Commons (CC)

As licenças CC são globais, legíveis por máquina e juridicamente válidas, garantindo a sua universalidade. Ao aplicar uma licença CC, o criador garante o crédito pelo seu trabalho e a titularidade dos direitos de autor (copyright ownership) permanece sempre com o licenciador, em contraste com a transferência de direitos para editores.

As licenças Creative Commons (CC) baseiam-se no princípio de que apenas alguns direitos são reservados:

Licença	Permite uso comercial?	Permite alterações?	Exige atribuição?
CC BY (Atribuição, *By*)	✅ Sim	✅ Sim	✅ Sim
CC BY-SA (Atribuição – Partilha Igual / *Share Alike*)	✅ Sim	✅ Sim, desde que com a mesma licença	✅ Sim
CC BY-ND (Atribuição – Sem Derivações / *No Derivatives*)	✅ Sim	❌ Não	✅ Sim
CC BY-NC (Atribuição – Não Comercial / *Non-Commercial*)	❌ Não	✅ Sim	✅ Sim
CC BY-NC-SA (Atribuição – Não Comercial – Partilha Igual / *Non-Commercial Share Alike*)	❌ Não	✅ Sim, desde que com a mesma licença	✅ Sim
CC BY-NC-ND (Atribuição – Não Comercial – Sem Derivações / *Non-Commercial No Derivatives*)	❌ Não	❌ Não	✅ Sim
CC0 (Domínio Público / *Public Domain Dedication*)	✅ Sim	✅ Sim	❌ Não

A CC BY é a licença mais permissiva, permitindo o uso máximo, incluindo para fins comerciais, desde que o crédito seja dado. A CC BY-NC-ND é a mais restritiva das licenças Creative Commons.

É importante notar que o uso de cláusulas restritivas como ND (No Derivatives - Proibição de Obras Derivadas) pode impedir a tradução de artigos para outras línguas, dificultando a disseminação global do conhecimento. Da mesma forma, a cláusula NC (Non-Commercial) pode criar ambiguidades jurídicas em blogs educativos ou plataformas que utilizem publicidade mínima para manutenção, pelo que a licença CC BY continua a ser a mais eficaz para garantir a máxima interoperabilidade e reutilização.

3.7. Mandatos de Licenciamento e Retenção de Direitos

Dados e Metadados: Dados abertos devem ser licenciados sob CC BY ou CC0 ou equivalentes, sendo esta a preferência para o Horizonte Europa (HE). Os Metadados de acesso aberto devem ser disponibilizados sob CC0 (domínio público), o que é um requisito da FCT e HE, e otimiza a reutilização por máquinas. O PGD deve especificar a licença a aplicar aos dados.

Publicações (Retenção de Direitos): Financiadores como a FCT e o Horizonte Europa exigem que publicações peer-reviewed estejam em Acesso Aberto imediato (sem embargo), sob licença CC BY 4.0 ou equivalente. A Estratégia de Retenção de Direitos (RRS) é utilizada para que o autor retenha os direitos sobre o Manuscrito Aceite do Autor (AAM) através de uma cláusula específica na submissão, garantindo que o publisher não ganhe a titularidade do AAM, e permitindo o depósito imediato (Via Verde) do AAM num repositório sob CC BY 4.0. O incumprimento destas regras resulta em perda de elegibilidade. Esta estratégia combate diretamente o modelo tradicional de "*Copyright Transfer Agreements*", onde os autores cediam todos os direitos às editoras. No contexto europeu, discute-se cada vez mais o direito de publicação secundária, que visa garantir por lei o direito de um investigador disponibilizar publicamente os seus resultados financiados pelo Estado, independentemente de contratos assinados com editoras comerciais.

Outros Resultados: Monografias e Capítulos de Livro podem usar licenças mais restritivas (CC BY-NC ou CC BY-ND). Teses e Dissertações (FCT) exigem uma licença CC com embargo máximo de 12 meses. O Software Aberto deve ter licenciamento (Ex: MIT, GPL, Apache) que conceda o direito de usar, aceder, modificar, expandir, estudar e partilhar o código fonte.

O PGD deve descrever a estratégia para disponibilizar as ferramentas e/ou softwares necessárias para o acesso, uso e reutilização dos dados.

3.8. Exemplo da Licença MIT

A Licença MIT é um exemplo de licença Open Source que permite que qualquer pessoa obtenha uma cópia do software e da documentação associada e lide com o software sem restrições. Isto inclui os direitos de usar, copiar, modificar, distribuir, sublicenciar e/ou

vender cópias, sujeito à inclusão do aviso de copyright e da permissão em todas as cópias ou porções substanciais. A licença declara que o software é fornecido "tal como está" ("*as is*"), sem garantia.

Copyright © 2025 <copyright holders>

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND,EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

3.9. Caso de Estudo: Escolha de Licença (Smart City)

Neste caso de estudo o projeto Smart City gera artigos (protegidos por DA), conjuntos de dados de sensores (protegidos pelo Direito Sui Generis) e software (protegido por DA).

Cenário 1: Publicação em Revista Híbrida

Para cumprir o mandato de AA imediato num projeto financiado, o autor deve aplicar a Estratégia de Retenção de Direitos no momento da submissão. Embora o autor possa pagar APCs (que não são elegíveis para financiamento) para a publicação na revista, a estratégia garante que o Manuscrito Aceite (AAM) seja depositado imediatamente (sem embargo, Via Verde) num repositório (como o Zenodo, que atribui DOI) sob a licença CC BY 4.0, cumprindo assim os requisitos de financiadores.

Estratégia de Retenção de Direitos (RRS): O aviso de Retenção de Direitos deve ser incluído no momento da submissão. A versão completa do aviso deve declarar que o manuscrito é baseado em resultados de investigação financiada por fundos públicos (ex.: FCT / Comissão Europeia) e que os autores retêm o direito de disponibilizar publicamente o AAM com uma licença CC BY 4.0, sem embargo, em repositórios reconhecidos (ex.: Repositório ULisboa, Zenodo), prevalecendo este aviso sobre quaisquer disposições contratuais em contrário.

Cenário 2: Licenciamento de Dados Não Sensíveis e Software

Dados: Conjuntos de dados não sensíveis (como qualidade do ar) devem ser publicados em CC BY 4.0 e os metadados em CC0 no Zenodo, destacando-se a ausência de restrições. A CC BY 4.0 ou CC0 é a escolha recomendada para maximizar a abertura.

Inclusão de Licença CC-BY: A licença deve ser clara e completa, incluindo o nome, a versão (ex.: 4.0) e o link oficial para o texto da licença.

Software: O software de recolha e organização de dados deve usar uma Licença Open Source (MIT) que permita a modificação e partilha do código fonte. O código deve ser partilhado no GitHub e ligado ao Zenodo (DOI). O software com potencial comercial (como o de visualização 3D) pode ter a sua publicação adiada por 12 meses para análise comercial.

Cenário 3: Dados Restritos e Controlo Para dados restritos

Para casos em que a anonimização robusta é impossível, por exemplo, vídeos de carros, onde pessoas são filmadas, aplica-se o princípio "Tão aberto quanto possível, tão fechado quanto necessário". A solução é o Acesso Controlado ("*managed access procedure*"), formalizado por um Acordo de Partilha de Dados (DSA) para autorizar a reutilização por terceiros sob condições restritas, salvaguardando obrigações do RGPD. Os metadados (título, autor, licença, restrições) devem ser CC0 e abertos.

Risco de Reidentificação: Um estudo de Latanya Sweeney demonstrou que, ao aplicar filtros a uma base de dados de registos de saúde supostamente anonimizados de funcionários públicos de Massachusetts, foi possível identificar o histórico médico do Governador William Weld. A maioria das pessoas nos EUA é a única no seu código postal com uma data de nascimento específica, tornando fácil descobrir as suas identidades. Este estudo, e diversos casos similares, reforçam a importância de fechar dados através de um DSA para mitigar riscos de reidentificação.

Cenário 4: Pre-prints de Artigos

A partilha de pre-prints é uma prática recomendada de Ciência Aberta para partilha antecipada, mas não cumpre os requisitos de Acesso Aberto obrigatórios dos financiadores, que exigem o depósito do manuscrito revisto por pares (AAM). O pre-print deve ser depositado num repositório confiável (não em redes sociais académicas) e o seu registo deve ser atualizado com o DOI da versão final. Deve ser usada uma licença cautelosa (não exclusiva) na versão inicial para não comprometer futuras publicações.

3.10. Obrigações e Boas Práticas

Obrigatório:

- É obrigatório fornecer Acesso Aberto imediato (sem embargo) a artigos científicos (AAM ou VoR) sob licença CC BY 4.0 ou equivalente ou o machine-readable electronic copy da versão publicada.
- O Acesso Aberto deve ser imediato (sem embargo).
- A publicação deve ser depositada num Repositório Confiável. Um repositório confiável deve ter mecanismos para garantir a precisão, integridade e autenticidade dos seus conteúdos.
- Os Metadados de dados de acesso aberto devem estar sob CC0.
- O PGD deve definir a propriedade dos dados (direitos de controlo de acesso), a gestão da PI e a estratégia para disponibilizar as ferramentas e/ou softwares necessárias para o acesso, uso e reutilização dos dados.

Boas práticas adicionais:

- Recomenda-se publicar em revistas Open Access (Gold OA), cujos custos são elegíveis. Os custos para publicações em revistas full Open Access (Gold OA) são elegíveis, mas as APCs para revistas híbridas não são elegíveis.
- A Estratégia de Retenção de Direitos deve ser utilizada preventivamente em todas as submissões a revistas.

- Licenças CC mais restritivas (CC BY-NC/ND) só devem ser usadas quando justificadas (ex: monografias ou para proteger interesses comerciais).
- Licenciar dados brutos (raw data) com CC0 ou CC BY resolve dúvidas sobre direitos conexos e maximiza a abertura.
- O Acesso Controlado para dados confidenciais ou sensíveis deve ser formalizado através de Acordos de Partilha de Dados (DSAs).
- O software de investigação deve ser partilhado sob licenças Open Source (Ex: GPL, MIT).
- O depósito em plataformas comerciais como ResearchGate, Academia.edu, websites pessoais ou serviços cloud (Dropbox, Google Drive) é explicitamente não conforme e nenhum destes é considerado um Repositório Confiável.

4. Regulamento Geral de Proteção de Dados (RGPD)

4.1. Conceitos Chave

O RGPD (Regulamento UE 2016/679 e Lei 58/2019 em Portugal) é o principal condicionante legal, garantindo que a Acessibilidade (A de FAIR) não viole a privacidade. O RGPD exige a implementação da proteção de dados desde a conceção e por padrão (*Data Protection by Design and Default*). As obrigações incluem p.ex. a minimização de dados, limitação da finalidade, e accountability (manter registos). O RGPD, considerado uma das legislações de privacidade de dados mais rigorosas do mundo, baseia-se em sete princípios centrais, incluindo a licitude, lealdade e transparência, limitação da finalidade, minimização dos dados, exatidão, limitação da conservação, integridade e confidencialidade, e responsabilidade (accountability).

Dados Pessoais cobrem qualquer informação sobre pessoas vivas que podem ser identificadas direta ou indiretamente, incluindo identificadores diretos (nome, email) e indiretos fortes (números de identificação, dados de geolocalização, endereço IP, ou fatores específicos da identidade física, psicológica, cultural e social). Dados Sensíveis (categorias especiais) incluem, por exemplo, dados genéticos, biométricos, ou relativos à saúde, à orientação sexual ou à raça

A Avaliação de Impacto (DPIA/AIPD) é obrigatória quando o tratamento de dados pessoais pode resultar num risco elevado para os direitos dos titulares dos dados, como em casos de dados sensíveis em larga escala ou utilização de novas tecnologias (p.ex: dados genéticos/saúde, user profiling).

A Anonimização Robusta retira os dados do âmbito do RGPD (permitindo a partilha aberta), enquanto a Pseudonimização é reversível (mantém os dados sob o RGPD). A pseudonimização é um processamento que permite a reidentificação através de informações adicionais armazenadas separadamente. Se a anonimização robusta não for possível, o acesso deve ser controlado (managed access) via Acordo de Partilha de Dados (DSA).

4.2. Aplicação do RGPD na Gestão de Dados Pessoais - Legislação Aplicável e Enquadramento Nacional

A legislação aplicável, na ordem jurídica nacional, que assegura a execução do RGPD é a Lei n.º 58/2019, a qual também regula o regime sancionatório aplicável às organizações, bem como o enquadramento penal. Outras leis relevantes incluem a Lei n.º 59/2019, que estabelece regras para o tratamento de dados pessoais para efeitos de prevenção, deteção, investigação ou repressão de infrações penais ou execução de sanções penais, transpondo a Diretiva (UE) 2016/680, e a Lei n.º 34/2009, que estabelece o regime jurídico aplicável ao tratamento de dados referentes ao sistema judicial.

Alguns diplomas secundários incluem a Portaria n.º 312-A/2022, que define as condições para transferências de dados pessoais para fora do Espaço Económico Europeu (EEE), e a Resolução do Conselho de Ministros n.º 41/2018, que define orientações técnicas de segurança para a Administração Pública.

Definição de Dados Pessoais e Sensíveis

Dados Pessoais englobam mais do que apenas identificadores diretos como o nome. Incluem números de identificação, dados de geolocalização, endereço IP, ou um ou mais fatores específicos da identidade física, psicológica, cultural e social de uma pessoa singular.

Dados Sensíveis são categorias especiais que incluem dados genéticos, biométricos ou relativos à saúde, bem como dados relativos à orientação sexual ou raça, e dados sobre condenações penais e infrações.

Projetos que envolvem participantes humanos, como ensaios clínicos, têm a obrigatoriedade de ter documentação ética. O RGPD é aplicável mesmo aos dados recolhidos do domínio público, como redes sociais.

Direitos dos Titulares dos Dados

O RGPD confere aos titulares dos dados os seguintes direitos:

- Direito de Acesso: Obter a confirmação de que os dados pessoais lhes dizem respeito e aceder à informação relativa aos dados que são tratados e às respetivas finalidades.

- Direito de Retificação: Retificar os dados pessoais ou retificar dados incompletos ou imprecisos a qualquer momento. O controlador tem a obrigação de corrigir dados incorretos ou incompletos.
- Direito ao Apagamento: Apagar os dados pessoais quando um prazo de conservação seja atingido ou o tratamento de dados deixe de ser lícito.
- Direito à Limitação do Tratamento: Suspender o tratamento dos dados pessoais.
- Direito à Portabilidade: Receber os dados pessoais num formato estruturado e eletrónico para transmitir a outra entidade.
- Direito de Oposição ao Tratamento: Opor-se ao tratamento, por motivos relacionados com uma situação particular válida que prevaleça sobre os direitos, liberdades e garantias.
- Direito à Oposição de Decisões Automatizadas: Não ficar sujeito a nenhuma decisão tomada exclusivamente com base no tratamento automatizado, incluindo a definição de perfis. Este direito de oposição é particularmente relevante, uma vez que o titular dos dados tem o direito de não ser sujeito a decisões tomadas exclusivamente com base no tratamento automatizado.
- Direito de Retirar o Consentimento: Retirar o consentimento do tratamento em específico a qualquer altura, por um motivo específico.

4.3. Bases Legais, Transparência e PGD

O tratamento de dados pessoais para arquivo e/ou investigação deve ter uma base legal. As bases legais mais comuns são o interesse público na investigação científica (para investigação profissional/doutoral) ou o consentimento explícito. Se o consentimento for a base legal, o participante tem o direito de o retirar a qualquer momento, o que obriga à eliminação dos dados.

Existe uma **Obrigação de Transparência**, segundo a qual todos os participantes têm o direito de saber como os dados serão processados, para que fins e a quais partes os dados serão transferidos. Esta informação deve ser fornecida numa Notificação de Privacidade (Privacy Notice) antes de a recolha de dados começar. A informação dada no Privacy

Notice é vinculativa, e qualquer alteração ao propósito deve ser notificada aos participantes. A Notificação de Privacidade deve ser concisa e redigida em linguagem clara e acessível, contendo obrigatoriamente a identidade e o contacto do controlador, a base legal e a finalidade do tratamento, as categorias de dados recolhidos, os direitos dos participantes e o tempo de retenção.

O Plano de Gestão de Dados (PGD) tem de explicar como será assegurado o cumprimento da legislação e regulamentação sobre dados pessoais (e o Encarregado de Proteção de Dados, DPO). Além disso, o tratamento de dados pessoais, como vídeo, tem de ser registado na Comissão Nacional de Proteção de Dados (CNPD) e/ou outros países no caso de projetos internacionais. No contexto de investigação, a organização que realize um projeto, como uma Universidade (ou o empregador) é o controlador dos dados, sendo responsável legal por garantir o cumprimento da legislação

4.4. Proteção de Dados desde a Conceção

O RGPD exige a implementação da proteção de dados desde a conceção e por padrão (*Data Protection by Design and Default*).

As Obrigações-Chave do RGPD são as seguintes:

1. Minimização de Dados: Recolher e reter apenas o que é estritamente necessário. A minimização de dados deve ser aplicada desde a fase inicial da recolha, removendo identificadores diretos ou indiretos fortes desnecessários. Deve-se evitar recolher dados desnecessários por mera dúvida, ou por omissão.

2. Limitação da Finalidade: Usar os dados apenas para o propósito específico para o qual foram recolhidos. Isto implica que os dados não podem ser usados para outros propósitos sem informar os participantes.

3. Integridade e Confidencialidade: Garantir a segurança e proteção dos dados contra acesso não autorizado. Os dados pessoais são considerados informação confidencial, e a obrigação de confidencialidade é ilimitada no tempo. A encriptação por si só pode ser insuficiente se o dispositivo de armazenamento estiver fisicamente acessível.

4. Accountability (Responsabilidade): Exige manter registos de pedidos e respostas.

O tempo de conservação dos dados pessoais deve ser definido e limitado, pois não podem ser mantidos indefinidamente.

No contexto da proteção de dados sensíveis, deve-se considerar ainda a geração de dados sintéticos (*Synthetic Data*) como alternativa à partilha de dados reais, permitindo que outros investigadores testem modelos estatísticos sem nunca acederem à informação original protegida pelo RGPD.

4.5. Avaliação de Impacto sobre a Proteção de Dados (DPIA)

A Avaliação de Impacto sobre a Proteção de Dados (AIPD/DPIA) é um documento interno, obrigatório para atividades de tratamento de dados de alto risco ou dados sensíveis em larga escala, e deve ser revisto pelo DPO.

Os exemplos de tratamento de alto risco incluem os seguintes casos:

- Criação de perfis (profiling) com base em tratamento automatizado com impacto legal, como pontuações de crédito.
- Tratamento de dados genéticos, biométricos e de saúde, ou dados relativos à orientação sexual ou raça.
- Tratamento de dados sobre condenações penais e infrações.
- Monitorização de zonas acessíveis ao público (câmaras de vigilância em larga escala).
- Utilização de novas tecnologias (como Inteligência Artificial / Machine Learning)
- Tratamento de dados de pessoas vulneráveis, como menores ou idosos, para fins de marketing.
- Aplicações móveis que rastreiam a localização de indivíduos.

As medidas de mitigação incluem salvaguardas técnicas e organizacionais para reduzir riscos. As Etapas da AIPD incluem:

1. Descrever o tratamento (Natureza, âmbito, finalidade e base legal);

2. Avaliar a necessidade e proporcionalidade;

3. Identificar riscos (acesso não autorizado, perda ou uso indevido de dados);

4. Definir medidas de mitigação;

5. Demonstrar conformidade, documentando o processo como prova de cumprimento do RGPD. A AIPD é um documento interno que pode ser exigido por uma comissão de ética em investigação.

4.6. Ciclo de Vida dos Dados Pessoais e Salvaguardas

Planeamento no PGD e Revisão Ética

As restrições éticas devem ser consideradas desde o início do processo de investigação. O PGD deve abordar explicitamente a proteção de dados pessoais.

A Revisão Ética é necessária para a recolha de dados que envolvam participantes humanos (por exemplo, por uma comissão de ética), e deve ser feita antes de começar a recolha de dados. A Revisão Ética é obrigatória se houver desvio do consentimento informado, intervenção na integridade física, pesquisa com menores/grupos vulneráveis ou exposição a estímulos fortes ou se forem recolhidos dados sensíveis.

Notificação de Privacidade e Consentimento

O Consentimento Informado deve ser livre, específico, informado e inequívoco. A participação deve ser voluntária, e os participantes devem ter o direito de se retirar a qualquer momento O formulário de consentimento deve cobrir a recolha, o uso e a futura partilha dos dados. No caso de menores, o consentimento dos pais é necessário, mas a criança deve ainda ter o direito de recusar a participação.

A Notificação de Privacidade (Privacy Notice) é obrigatória antes da recolha de dados. É uma informação vinculativa; se for atualizada (por exemplo, sobre novos fins), os participantes devem ser notificados. Recomenda-se o uso de templates institucionais e atualizados.

Os elementos obrigatórios do Privacy Notice incluem:

1. Identidade e contacto do controlador dos dados (instituição).

2. Bases legais e finalidades do tratamento.
3. Categorias de dados recolhidos.
4. Direitos dos participantes (acesso, remoção, etc.).
5. Transferências fora da UE.
6. Detalhes do DPO.
7. Tempo de retenção.

Armazenamento e Segurança de Dados Confidenciais

O armazenamento deve ser feito em sistemas em rede seguros. Deve ser evitado o armazenamento em pen drives, discos rígidos locais ou cloud públicas não autorizadas, sendo recomendados serviços cloud institucionais localizados no Espaço Económico Europeu (EEE).

O armazenamento deve ser feito em sistemas em rede seguros. O PGD deve descrever os procedimentos de backup (por exemplo, para cloud e tape backups). Dados de investigação, na prática, geralmente têm o nível de "confidencial" e devem ser encriptados. Dados classificados como "secretos" (como registos de pacientes) requerem ambientes isolados de alta segurança (p.ex., SECdata) e não podem ser descarregados ou carregados sem intervenção administrativa.

A gestão de dados classificados (Dual Use) requer ambientes de armazenamento seguros, que podem incluir restrições de acesso físico e encriptação.

Anonimização vs. Pseudonimização e Preservação

A Anonimização Robusta torna a reidentificação impossível, colocando os dados fora do âmbito do RGPD. A Pseudonimização é reversível (por exemplo, encriptação) e os dados continuam cobertos pelo RGPD. A ferramenta AMNESIA do OpenAIRE é um recurso open source que pode ser usado livremente para a anonimização de dados sensíveis.

Os dados pessoais só podem ser arquivados a longo prazo após anonimização ou pseudonimização. A transferência de dados pessoais para países fora do Espaço Económico Europeu (EEE) deve ser comunicada aos participantes e requer conformidade rigorosa com o RGPD. A transferência para fora do EEE exige medidas especiais devido ao

risco de o RGPD não ser aplicado.

Acesso Controlado

A restrição de dados é válida quando a acessibilidade (A de FAIR) não pode violar a privacidade. Se a anonimização robusta for impossível (por exemplo, vídeos de linguagem gestual ou de trânsito), a solução é o Acesso Controlado (managed access procedure). Isto exige um protocolo de autenticação e autorização de utilizadores, regulado por um Acordo de Partilha de Dados (DSA), e o registo de acesso (logs) deve estar previsto no PGD.

Obrigações no PGD (Modelo FCT)

O PGD deve demonstrar a conformidade legal. O modelo PGD da FCT exige o seguinte:

1. Dados Pessoais: Confirmar se serão processados e como será assegurado o cumprimento da legislação.
2. Segurança: Descrever as medidas de segurança adicionais aplicadas a dados sensíveis.
3. PI/Propriedade: Gerir a propriedade dos dados e os direitos de Propriedade Intelectual (PI),incluindo trade secrets.

4.7. Caso de Estudo: Anonimização e Consentimento

Cenário: Dados Pessoais em Entrevistas

Um projeto de investigação recolhe entrevistas com dados de geolocalização e socioeconómicos (dados sensíveis e pessoais). A questão ética é garantir a partilha futura (R de FAIR) e a partilha internacional sem violar o RGPD. A documentação ética é obrigatória para participantes humanos, e a Revisão Ética é necessária antes de iniciar a recolha de dados.

Aplicação do Consentimento e Privacidade (Exemplo)

1. Privacy Notice: Informar o participante sobre o propósito, os dados recolhidos, a eventual transferência de dados fora da UE e os propósitos de reutilização. Se o estudo envolver comunidades vulneráveis, os Princípios CARE devem ser considerados para garantir que a comunidade tem autoridade para controlar o acesso (Authority to Control) e beneficia coletivamente do estudo.

2. Consentimento: Deve ser explícito e inequívoco para a partilha futura (mesmo que seja anonimizada ou restrita).

Anonimização Prática e Partilha Condicional

A solução técnica preferencial é a aplicação de anonimização robusta (por exemplo, agregação de dados, supressão de datas exatas e de geolocalização individual) para remover os dados do âmbito do RGPD e permitir a partilha aberta e internacional. Ferramentas como o AMNESIA podem ser usadas.

Se a anonimização não for possível, a Partilha Condicional é a solução: o acesso controlado é gerido via DSA e os dados devem ser mantidos em ambientes seguros (encerrados).

Exemplo de Aviso de Privacidade RGPD (Versão Curta): *Os dados recolhidos destinam-se exclusivamente a fins de investigação científica e serão tratados de forma confidencial e anonimizados antes de qualquer partilha ou publicação. O participante pode, a qualquer momento, solicitar acesso, retificação, eliminação ou retirar o consentimento. O tratamento cumpre o RGPD e a Lei 58/2019.*

Exemplo de Aviso de Privacidade RGPD (Versão Longa): *A responsabilidade do tratamento é da instituição, sendo a finalidade exclusiva de investigação científica no âmbito do projeto, analisando objetivos específicos (ex.: padrões socioeconómicos e de mobilidade urbana). Serão recolhidos dados pessoais e sensíveis, sendo que, sempre que possível, os dados serão anonimizados ou pseudonimizados. A partilha e reutilização de dados para fins científicos será efetuada apenas após anonimização robusta ou, se esta não for possível, o acesso será restrito e controlado mediante Acordo de Partilha de Dados (DSA) em ambientes seguros. A base legal é o consentimento explícito, que pode ser retirado a qualquer momento. Os dados pessoais identificáveis serão conservados apenas pelo tempo necessário à finalidade da investigação e eliminados ou anonimizados após o prazo definido na aprovação ética.*

4.8. Obrigações e Boas Práticas

Obrigatório:

- Cumprir a legislação sobre dados pessoais (RGPD) e garantir a proteção e segurança dos dados sensíveis, integrando as questões de proteção de dados no

PGD.

- O processo de anonimização ou pseudonimização aplicado aos dados deve ser explicitamente detalhado no PGD.
- Obter consentimento informado para a recolha, preservação e/ou partilha de dados pessoais.
- Fornecer uma Notificação de Privacidade (Privacy Notice) clara antes de começar a recolha de dados.
- Garantir Backup e armazenamento seguro, encriptando dados confidenciais.
- Realizar uma Avaliação de Impacto sobre a Proteção de Dados (DPIA/AIPD) para o tratamento de dados de alto risco, com riscos e as medidas de mitigação planeadas
- Manter registos de pedidos e respostas (princípio da accountability).

Boas práticas adicionais:

- Incluir todos os direitos dos titulares nos formulários de consentimento e na documentação de privacidade.
- Envolver o Encarregado de Proteção de Dados (DPO) da instituição em todas as fases do projeto.
- Garantir anonimização ou pseudonimização sempre que possível para reduzir riscos.
- Seguir os códigos de conduta nacionais e internacionais e diretrizes éticas institucionais.
- Usar templates institucionais para o Privacy Notice e Consentimento, ou templates para um procedimento de acesso controlado (managed access procedure) e Acordos de Partilha de Dados (DSAs) para dados restritos.
- Usar ferramentas de anonimização.

5. Normas Internacionais e Boas Práticas

5.1. Conceitos Chave

Este guia aborda os padrões de qualidade para garantir a Interoperabilidade (I de FAIR). Isto inclui o uso de Identificadores Persistentes (PIDs), como DOI para dados, ORCID para autores, ROR (*Research Organization Registry*) para organizações, SWHID para software e o Crossref Funder ID para a identificação do financiador. O PID Graph é recomendado para Scientific Knowledge Graphs (dados complexos), atribuindo um URI adicional para agregar todos os recursos digitais de um projeto.

Esta rede de identificadores permite o rastreio automatizado do impacto da investigação e a atribuição correta de créditos. Tal é essencial para garantir a rastreabilidade e a citação estável, sendo os PIDs fundamentais para o Findable (F de FAIR), garantindo que os dados sejam citáveis e rastreáveis.

Os Metadados (a descrição dos dados, ou as etiquetas dos dados) devem ser padronizados, p.ex. o Dublin Core (15 elementos, universal), e standards específicos de domínio (DICOM, ISO 19115). O Fairsharing.org é um catálogo de referência para estes padrões, e para identificar padrões de metadados disciplinares específicos. A interoperabilidade exige que se use formatos sem perda (lossless).

A adoção de maDMPs (Machine-actionable DMPs) PGDs legíveis por máquina, e o uso de formatos abertos (ex: CSV) são cruciais para a automação e preservação.

5.2. Standards Internacionais para a Gestão de Dados

PIDs (Identificadores Persistentes)

Os Identificadores Persistentes (PIDs), como DOI, ORCID e ROR, têm a função principal de estabelecer ligações estáveis entre todas as entidades da investigação: investigadores, dados, software, publicações e instituições. Os PIDs permitem a identificação inequívoca de cada entidade, mesmo quando ocorrem mudanças (nomes, afiliações, URLs).

Os PIDs suportam a agregação automática de publicações (ex: no Google Scholar, Scopus,

Web of Science) e a agregação e análise de citações em serviços bibliométricos. São fundamentais para melhorar a visibilidade científica, a interoperabilidade e a rastreabilidade dos outputs, contribuindo diretamente para o princípio Localizável (F) do FAIR.

Tipos de PIDs e suas funções:

- DOI: Utilizado para dados, publicações e software.
- ORCID: Identifica autores, sendo obtido em orcid.org.
- ROR: Identifica instituições.
- RAID: Utilizado para projetos e atividades.
- SWHID (Software Heritage Identifier): Identificador específico para software (As soluções para dados podem não ser adequadas para software executável).
- PID Graph: É recomendado para Scientific Knowledge Graphs (dados complexos), atribuindo um URI adicional para agregar todos os recursos digitais de um projeto.

Tipo de PID	Para quê?	Quem gere?	Onde registar / obter?
DOI	Dados, publicações, software	P.ex. DataCite (para dados) / CrossRef (para publicações)	P.ex. Zenodo, Repositório ULisboa
ORCID	Identificar autores	ORCID Inc.	orcid.org
ROR	Identificar instituições	ROR Community	ror.org
RAID	Projetos / atividades	ARDC	raid.org
ISBN/ISSN	Livros / Revistas	Agências ISBN/ISSN	P.ex. Biblioteca Nacional
Handle	Objetos digitais	CNRI / DONA foundation	handle.net
ARK	Preservação digital	California Digital Library	arks.org

Proveniência e Versionamento

A rastreabilidade e a proveniência (provenance) referem-se ao histórico de versões e alterações de um objeto digital. A boa prática exige que o Plano de Gestão de Dados (PGD) defina se uma alteração menor (ex: correção ortográfica) deve ser apenas registada no registo de alterações (change log) do PID existente, ou se uma alteração maior (ex: mudança de um dataset) deve despoletar a criação de um novo PID.

Formatos de Ficheiro

A escolha de formatos abertos é crucial para a interoperabilidade e a preservação a longo prazo, uma vez que formatos proprietários (como .XLSX, .DOCX, .SPSS) dependem de licenças pagas, têm interoperabilidade limitada e maior risco de obsolescência. Os formatos proprietários dependem de licenças pagas e das versões, têm interoperabilidade limitada e maior risco de obsolescência, dificultando a leitura em softwares alternativos e comprometendo a preservação a longo prazo.

Categoria	Formato	Tipo / uso e software	Justificação para uso dos formatos abertos
Proprietário	.XLSX	Folha de cálculo (Microsoft Excel)	Funciona bem em software específico, mas depende de licença paga e das versões; interoperabilidade é limitada com outros softwares; risco maior de obsolescência.
Proprietário	.DOC / .DOCX	Documento de texto (Microsoft Word)	Formato fechado, diferentes versões podem ter problemas de compatibilidade; difícil de garantir acesso futuro sem software Microsoft.
Proprietário	.SPSS	Dados estatísticos (IBM SPSS)	Proprietário e binário; dificulta leitura em softwares alternativos; preservação de longo prazo comprometida.
Aberto	.CSV	Dados tabulares	Formato simples e universal; facilmente lido por qualquer software; ideal para interoperabilidade e preservação.
Aberto	.PDF/A	Documento (versão para arquivo do PDF)	Formato aberto e padronizado (*ISO-standardized*); preservação a longo prazo garantida; mantém layout e integridade do documento.
Aberto	.ODT	Documento de texto (OpenDocument)	Aberto e padronizado; suporta interoperabilidade entre diferentes software *office suites*; fácil acesso futuro.
Aberto	.TIFF	Imagem	Formato de imagem aberto ideal para preservação a longo prazo de documentos digitalizados; amplamente suportado.

Vocabulários Controlados e Ontologias

Para assegurar a Interoperabilidade (I de FAIR) e o processamento semântico por máquinas são usados os vocabulários controlados e as ontologias.

Vocabulário Controlado: É uma lista de termos padronizados para garantir a consistência na comunicação e na recuperação de informações, controlando a terminologia para

evitar ambiguidades. O vocabulário controlado assegura consistência, precisão e clareza na comunicação e na catalogação de informações, e facilita a busca e recuperação de informações.

Ontologia: É uma representação formal e estruturada de conhecimento que define não só os termos, mas também as relações e hierarquias entre eles. Uma ontologia, como um modelo de saúde que liga doenças, sintomas e tratamentos, modela o conhecimento de uma área para que possa ser usado por máquinas para inferência e análise automática.

Standards de Metadados

Os metadados são a descrição dos dados e são essenciais para a Localização (F) e a Interoperabilidade (I). Podem ser Universais ou Específicos de Domínio.

Dublin Core (Metadados Universais): É um conjunto de 15 elementos simples, essenciais para descrever qualquer recurso digital (dados, publicações, software). É a linguagem base que todos os repositórios usam e é uma norma internacional adotada pela ISO. Estrutura: Concentra-se em Conteúdo (Título, Tópico, Tipo, Cobertura), Propriedade Intelectual (Criador, Direitos, Editor) e Instanciação (Data, Formato, Identifier – que deve ser um PID). Sem elementos como Título, Criador e Data, o recurso é invisível.

Metadados Específicos de Domínio: Descrevem o contexto científico e técnico do dataset (métodos, instrumentos, parâmetros). Enquanto o Dublin Core responde a "O que é o recurso?", os metadados de domínio respondem a "Como, onde e sob que condições foi gerado o recurso?".

Exemplos de Metadados Específicos de Domínio:

- DICOM: Para Imagiologia Médica, garante que a imagem inclui dados clínicos e técnicos de uso seguro (ex: PatientName e StudyDate/Time).
- ISO 19115: Para Geografia e Ciências da Terra, descreve a informação geográfica digital, incluindo a Geographic Bounding Box (coordenadas) e a Spatial Resolution.
- O Darwin Core (DwC) é um standard importante para padronizar informações sobre ocorrências de espécies, e é importante para dados de museus e coleções biológicas.
- Standards de domínio podem ser encontrados no Fairsharing.org.

Standards para Outros Resultados de Investigação

Repositórios Confiáveis: A CoreTrustSeal é uma certificação para repositórios que garante elevados padrões de confiança, preservação a longo prazo e interoperabilidade com a EOSC.

Software: O software de investigação exige standards próprios, como o SWHID (Software Heritage Identifier) para identificação e o CodeMeta como esquema de metadados para software.

Protocolos e Metodologias: Protocol.io permite a criação, partilha e atribuição de DOIs a protocolos detalhados, aumentando o sucesso da replicação de experiências.

Workflows Computacionais: WorkflowHub é usado para registar e partilhar workflows. O WorkflowHub visa ser o principal catálogo e centro de partilha para todos os workflows científicos computacionais publicamente disponíveis. Outro exemplo é o RO-Crate (Research Object Crate) agrupa datasets, software e workflows relacionados numa única estrutura de dados interligada, facilitando a sua transferência.

5.3. Frameworks e Boas Práticas na Ciência Aberta

O **Framework FAIR** é a estrutura principal de qualidade na Ciência Aberta:

- Findable (Localizável): PID (DOI), metadados ricos e possibilidade de pesquisa online.
- Accessible (Acessível): Depositado num trusted repository (e.g., Zenodo), seguindo o princípio "tão aberto quanto possível, tão fechado quanto necessário".
- Interoperable (Interoperável): Uso de formatos abertos (e.g., CSV em vez de XLSX) e termos padronizados.
- Reusable (Reutilizável): Bem documentado (e.g., README files), incluindo proveniência e ferramentas necessárias, e licença clara (e.g., CC BY 4.0, CC0). A acessibilidade é garantida ao saber onde os dados estão e como lhes aceder (protocolo).
-

Machine-Actionable DMP (maDMP)

O maDMP representa a evolução para formatos dinâmicos de PGDs legíveis por computador (JSON/XML). É dinâmico e permite a automação de tarefas de gestão de dados, como o pré- preenchimento de metadados no repositório, poupando tempo e permitindo a verificação automática de conformidade pelos financiadores. O ARGOS é a ferramenta recomendada que permite que o repositório leia o maDMP e preencha automaticamente os metadados.

A adoção de Planos de Gestão de Dados "Actionable" (legíveis por máquina) permite que a informação sobre os dados seja trocada automaticamente entre sistemas (ex: repositórios e financiadores). Ao criar um PGD, deve-se utilizar vocabulários controlados e ontologias específicas da área de estudo para garantir que as máquinas possam interpretar o contexto dos dados sem intervenção humana constante.

Organização de Dados e Nomeação de Ficheiros

É recomendada uma estrutura de pastas simples (flat structure), organizada por Fases do Projeto ou Tipo de Dados. Devem ser evitadas as sub-pastas aninhadas que dificultam a localização. O formato de data recomendado na convenção de nomeação é o ISO 8601 (AAAA-MM-DD).

- Pastas Recomendadas: 01_Raw_Data, 02_Processed_Data, 03_Scripts_Code, 04_Documentation.
- Convenção de Nomeação: Segue o princípio do geral para o específico, utilizando o formato de data ISSO 8601 (AAAA-MM-DD) (Ex: 20250521_ULisboa_WP3_EXP_Rato_02_v01_DadosBrutos_pH.csv).

Segurança: Dados confidenciais/sensíveis (RGPD) nunca devem ser armazenados em discos externos, pen drives ou cloud públicas não autorizadas. Deve ser usado armazenamento seguro na rede da instituição para garantir controlo de acesso, encriptação e auditoria.

Data Availability Statement (DAS)

A DAS é obrigatória em todos os artigos. É um elemento de transparência que garante que o dataset é Findable e Accessible. A DAS deve ser adicionada no final do artigo, antes da submissão. A DAS deve informar explicitamente a localização e as condições de acesso e

não deve instruir os leitores a contactar o autor para obter os dados (pois isso é uma falha de acessibilidade).

Workflows e Signposting

Workflows: Os workflows computacionais devem ser registados em repositórios (ex: WorkflowHub) e serem citados como produtos académicos formais. O uso do SWHID no workflow reforça a integridade e reflete modificações no código-fonte.

Signposting (Referenciação Precisa): Permite a navegação precisa de uma landing page para uma parte específica de um recurso, utilizando cabeçalhos HTTP da página para ligar o dataset, os metadados, a licença e a publicação associada. Isto torna as referências mais precisas e aumenta a reprodutibilidade.

Exemplo: rel="describedby" informa as máquinas onde recolher os metadados para indexação (F); rel="license" determina os termos de reutilização (R); e rel="isreferencedby" estabelece uma ligação bidirecional entre o dataset e a publicação associada.

Recursos Educativos Abertos (REA)

Salienta-se ainda que a eficácia dos recursos educativos abertos (REA) assenta no princípio dos 5 R's: Retenção (direito de fazer cópias), Reutilização (usar no formato original), Revisão (adaptar ou modificar), Remix (combinar com outros materiais) e Redistribuição (partilhar com outros). Ao criar REA, os autores devem garantir que o formato do ficheiro é editável (ex: .odt em vez de apenas .pdf) para que outros educadores possam efetivamente exercer estes direitos de adaptação pedagógica.

Monitorização de Impacto

No contexto da monitorização de impacto, plataformas como a The Lens oferecem dashboards de visualização e ferramentas de análise que permitem cruzar registos académicos com patentes globais (provenientes de 106 jurisdições). Esta integração é fundamental para demonstrar a transferência de conhecimento da investigação fundamental para a inovação tecnológica e industrial.

O Papel dos Data Stewards (Gestores de Dados)

Os Data Stewards dão apoio especializado em três áreas:

1. Conformidade: Aconselham sobre RGPD/Ética (anonimização, consentimento), escolha da licença e revisão de PGDs (garantindo o formato maDMP).

2. Qualidade: Aconselham na escolha de metadados específicos de domínio (ex: DICOM), indicam o repositório adequado (certificado) e promovem ferramentas para o registo da proveniência.

3. Treino: Capacitam investigadores sobre boas práticas de GDI (convenções de nomeação de ficheiros, standards) e garantem que os dados e os workflows são corretamente documentados.

5.4. Casos de Estudo (Exemplos de Aplicação de Standards)

Um projeto de Ecologia pode utilizar o standard EML (Ecological Metadata Language). Elementos como TaxonomicClassification usam vocabulários controlados para garantir a Interoperabilidade; Unit define a unidade de medida exata (Reutilizável); e GeographicCoverage define as coordenadas geográficas (Localizável). A prática de Ciência Cidadã (onde o público recolhe dados sobre espécies) pode ser aplicada, sendo que o standard Darwin Core (DwC) é importante para padronizar estas informações.

Em projetos de Smart City, para além dos standards de metadados geográficos (ISO 19115), são usados formatos abertos como GeoJSON para representar geometrias geográficas. Em termos de workflows (como modelos de previsão de tráfego), a equipa deve utilizar um fluxo de trabalho científico computacional registado no WorkflowHub e descrito no padrão CWL para garantir a portabilidade e a rastreabilidade.

5.5. Obrigações e Boas Práticas

Obrigatório: A Organização e o Investigador têm responsabilidades específicas. A Organização deve garantir a infraestrutura para emitir PIDs (DOI, Handle) e fornecer armazenamento seguro na rede.

- É obrigatório fornecer informação sobre softwares, algoritmos, protocolos,

modelos, workflows e electronic notebooks necessários para a validação das conclusões

- O Investigador deve obter o ORCID (obrigatório), aplicar o Dublin Core mínimo obrigatório e standards de domínio, adotar formatos abertos (CSV, PDF/A), usar convenções de nomeação com data ISO e redigir um README.md com a documentação.

- Publicar só em trusted repositories. Os repositórios confiáveis (Trusted Repositories) devem ter serviços, mecanismos e provisões para assegurar a precisão, integridade, autenticidade e acesso dos conteúdos.

Boas práticas adicionais:

- Utilizar ferramentas que suportam o formato maDMP e licenciar dados abertos com CC BY 4.0 ou CC0.

- Plano de Gestão de Dados (PGD) - criação de um modelo de maDMPnuma ferramenta (p.ex. ARGOS ou OpenAIRE).

- Metadados & Standards - Adoção de standards de domínio (ex: Darwin Core, EML, SWHID) através dos Data Stewards.

- Reutilização & Formato - Depósito de dados e software em formatos abertos (ex: CSV, PDF/A) para garantir a preservação.

- Recomenda-se utilizar ferramentas PGD que suportem a automação (maDMPs), como o ARGOS.

- A Data Availability Statement (DAS) deve ser adicionada no final do artigo antes da submissão.

- Apoio e Formação - Rede de Data Stewards e Data Champions para apoio especializado em GDI e RGPD.

6. Ferramentas e Recursos Disponíveis

6.1. Conceitos Chave

As ferramentas digitais são cruciais para a gestão e a reprodutibilidade. Este guia mapeia algumas ferramentas essenciais ao longo do ciclo de investigação, e para a gestão de dados. Existem, por exemplo, ferramentas de planeamento (ARGOS - PGDs maDMPs, OpenAIRE Costing Tool), de processamento e análise de dados (AMNESIA - anonimização, Jupyter Notebooks - proveniência, GitHub - versioning), e de publicação/preservação (Zenodo, Repositórios Certificados, re3data). O OpenAIRE funciona como o ecossistema central, liga repositórios e projetos (OpenAIRE Graph), e oferece serviços de valor acrescentado (ARGOS, AMNESIA, EXPLORE), que suportam a EOSC (European Open Science Cloud). O OpenAIRE Graph é um dataset aberto e abrangente de informação de investigação, cobrindo 166 milhões de publicações, 59 milhões de dados de investigação e 203 mil itens de software de investigação, de 131 mil fontes de dados, ligados a 3 milhões de bolsas e 193 mil organizações. O Graph permite ligar e visualizar todos os outputs da sua investigação – publicações, datasets, ORCID, software, PGD – interligados através de citações e semântica.

O ARGOS (OpenAIRE) é a ferramenta recomendada para criar PGDs machine-actionable (maDMPs) e inclui o modelo FCT (Anexo II). É um serviço gratuito e open source, configurável e extensível, para planear atividades de Gestão de Dados de Investigação (GDI) em conformidade com as políticas de Acesso Aberto (AA) e dados FAIR. O ARGOS está alinhado com o Research Data Alliance DMP Common Standard. O DMPonline é outra plataforma online de PGD. O Zenodo é o repositório genérico do CERN, que suporta uploads até 50 GB e se integra com o GitHub. O OpenAIRE atua como uma infraestrutura que conecta outputs de investigação (publicações, projetos, datasets) e oferece apoio através do Open Science Helpdesk e NOADs (National Open Access Desks).

No âmbito dos modelos de publicação, devem distinguir-se claramente as seguintes nuances:

Via de Diamante (Diamond OA): Considerada a forma mais pura de Acesso Aberto, onde não existem custos nem para o autor (sem APCs) nem para o leitor. É geralmente apoiada por instituições académicas ou sociedades científicas que valorizam o conhecimento como um bem comum.

Revistas Híbridas: Revistas sob subscrição que permitem tornar artigos individuais

abertos mediante pagamento. Deve-se ter cautela com o "Double Dipping", onde as bibliotecas institucionais pagam a subscrição da revista e os investigadores pagam adicionalmente para abrir os seus artigos na mesma publicação.

Acordos Transformativos (Transformative Agreements): São contratos "Read & Publish" negociados entre consórcios de bibliotecas e editoras, visando converter os orçamentos de subscrição em fundos para apoiar o custo de publicação em acesso aberto para todos os autores da instituição.

6.2. Repositórios e Plataformas para Gestão de Dados

As ferramentas e recursos disponíveis para GDI podem ser categorizados por fase do ciclo de investigação.

Fase do Ciclo de Investigação	Ferramentas Recursos (Exemplos)	Tipo de Ferramenta
1. Planeamento da investigação e desenho das fases do projeto	ARGOS (OpenAIRE), DMPonline, DS Wizard	Elaboração e gestão de PGDs machine-actionable
	OpenAIRE Costing Tool	Estimativa de custos de GDI/RDM
	Fairsharing.org	Busca e referência de standards e vocabulários de metadados
2. Recolha & processamento de dados	AMNESIA (OpenAIRE)	Anonimização e pseudonimização de dados sensíveis
	VeraCrypt ou 7-zip	Encriptação e armazenamento seguro de dados confidenciais
	GitHub/GitLab	Versionamento e documentação de código/software (open source)
	Jupyter Notebooks / eNotebooks	Documentação de procedimentos, análises interativas e registos de proveniência
3. Análise e elaboração de relatórios de investigação	Binder (mybinder.org)	Criação de ambientes computacionais reproduzíveis (código + dados)
	Open Research Europe (ORE)	Plataforma de publicação Open Access imediato (HE), que suporta artigos, protocolos e software
	NotebookLM / ChatPDF	Ferramentas de IA para interação e síntese de documentos/literatura
4. Publicação, partilha e preservação	Zenodo (OpenAIRE), Repositórios Institucionais (Certificados)	Depósito de dados, software e outros outputs com PIDs
	Re3data	Registo global para encontrar repositórios confiáveis
	OpenAIRE EXPLORE	Descoberta de pesquisa, ligações (*linking*) e monitorização de outputs

Sistemas e Repositórios de Ciência Aberta

Os sistemas de Ciência Aberta variam em escala e conteúdo, desempenhando papéis distintos na European Open Science Cloud (EOSC).

Tipo de Sistema	Escala	Conteúdos	Exemplos	Papel na EOSC
Institucional	Local	Publicações, dados, teses	Apollo, Aaltodoc, Repositório ULisboa	Fornece dados à rede
Temático	Europeu/ global	Dados normalizados por área	Zenodo, PANGAEA, ENA	Especialização disciplinar
Nacional	Nacional	Dados e publicações agregadas	HAL, RCAAP, NORA	Coordenação e política nacional
Pan-Europeu	Europeu	Serviços federados de dados	EOSC, OpenAIRE	Interoperabilidade continental
Software	Global	Código, scripts, pipelines	GitHub, Software Heritage	Reprodutibilidade
Apoio e Interoperabilidade	Institucional /global	Metadados e identificadores	PTCRIS, ORCID, DOI	Gestão e rastreabilidade

O Repositório Científico de Acesso Aberto da ULisboa assegura a interoperabilidade e exporta informação para o RCAAP (Repositórios Científicos de Acesso Aberto de Portugal), OpenAIRE e CIÊNCIAVITAE. Este sistema atua como o sistema principal onde os dados devem ser inicialmente carregados.

O Repositório Zenodo, mantido pelo CERN, é um repositório genérico crucial para datasets, software e outros outputs que não se enquadram em repositórios disciplinares. Atribui DOI automaticamente e permite uploads até 50 GB por registo. As suas funcionalidades incluem perfil de utilizador, depósito de publicações, com versões, revisão e colaboração, comunidades, e ligação com GitHub, ORCID e OpenAire. O processo de depósito no Zenodo envolve criar a conta e ligá- la ao ORCID, colocar ficheiros num novo registo, descrever o conjunto de dados, definir a visibilidade, e publicar ou partilhar acesso a drafts ou registos de acesso restrito.

Outros Repositórios incluem repositórios institucionais como ACRIS (Aalto University), RepositoriUM (U. Minho), Dspace@MIT, e Apollo (U. Cambridge), e repositórios de domínio como GenBank (Genética/Biologia Molecular), Pangaea (geociências e ambiente), e ICPSR (Ciências Sociais e Humanidades). O Global Re3data é um registo que ajuda investigadores a encontrar repositórios de dados confiáveis (trusted repositories), permitindo a pesquisa por repositórios registados no OpenAIRE ou que são certificados.

O Ecossistema OpenAIRE e a EOSC

O OpenAIRE é uma rede descentralizada e um ecossistema que liga vários repositórios e sistemas de informação através de diretrizes bem definidas, consolidando toda a informação num único Grafo Semântico (OpenAIRE Graph). O Graph recolhe e interliga metadados de mais de 70.000 fontes académicas em todo o mundo, ligando publicações a datasets, a software, a financiamentos, a organizações e a investigadores.

Os serviços de valor acrescentado do OpenAIRE incluem:

- OpenAIRE EXPLORE: Plataforma agregadora que liga publicações, datasets, software, PGDs e financiamentos, permitindo a descoberta de todos os outputs e projetos ligados no Graph. Pode ser usado para pesquisar um repositório para publicação.
- AMNESIA: Serviço de anonimização de dados pessoais.
- ARGOS: Ferramenta para criação de Planos de Gestão de Dados (PGDs) legíveis por máquina (maDMPs).
- OpenAIRE Broker: Enriquece os metadados dos repositórios com informação do Graph (ex: ligação a financiamentos, verificação de duplicados).
- UsageCounts e MONITOR: Permitem o acompanhamento das métricas de produção e conformidade por financiadores, instituições e gestores.
- OpenAIRE CONNECT e PROVIDE: Apoiam Comunidades de Investigação Específicas e conectam repositórios nacionais, garantindo a agregação e normalização de metadados.

O OpenAIRE é um prestador de serviços essenciais para a EOSC (European Open Science Cloud), atuando como o hub de interoperabilidade entre as diferentes fontes de dados. A EOSC Association (EOSC-A) é a entidade legal estabelecida para governar a European Open Science Cloud (EOSC) e foi formada em julho de 2020 com quatro membros fundadores, contando com mais de 250 membros e observadores. A EOSC é composta pelo EOSC Core (que implementa a estrutura de interoperabilidade, catálogo, marketplace, sistema PID e helpdesk) e o EOSC Exchange (onde os dados e serviços são partilhados através de protocolos de acesso estabelecidos). A Comissão Europeia está a criar 9 espaços de dados, sendo a EOSC o espaço dedicado à ciência, investigação e inovação.

Ferramentas para Planos de Gestão de Dados (PGD) e RDM

Identificadores Persistentes e Documentação Pessoal

O registo no ORCID é crucial para identificar investigadores. A informação necessária para o registo é nome, email e afiliação (opcional), e os dados são registados nos EUA. Após a atribuição, um link pessoal fica imediatamente disponível. É possível adicionar dados pessoais, websites, perfis de redes sociais, educação e qualificações, atividades

profissionais, financiamentos e trabalhos de investigação.

Elaboração e Gestão de PGDs

As ferramentas DMP online facilitam a preparação, geração e atualização de PGDs online, simplificando o trabalho dos investigadores e auxiliando no cumprimento dos requisitos de Ciência Aberta e de financiadores. Estas ferramentas incluem, por exemplo, ARGOS, DMPTuuli, DMP Opidor, DMPonline e EasyDMP.

Funcionalidades: Oferecem, em geral, templates, guias e uma lista de PGDs públicos. O DS Wizard, por exemplo, tem foco em Data Stewardship, integrações, guias FAIR e modelos de conhecimento para múltiplos perfis.

Custos de Gestão de Dados (RDM)

O OpenAIRE Costing tool é uma checklist que ajuda a determinar os custos da gestão de dados que devem ser incluídos no orçamento do projeto e no PGD (por exemplo, custos de staff, armazenamento, recolha de dados e curadoria). O DS Wizard também possui uma ferramenta similar, mais simples, que faz estimativas de custos, por exemplo, a 10 anos, com opções de configuração para unidades de armazenamento, backup, instalação e resposta a incidentes.

Standards de Metadados e Qualidade

O Fairsharing.org é um recurso para encontrar os standards e vocabulários controlados, bases de dados e políticas específicas para a área do domínio de investigação. O CEDAR workbench é uma ferramenta de apoio para a criação de metadados legíveis por máquinas (machine-readable) que utiliza templates de metadata para experiências biomédicas.

A checklist Think. Check. Submit. é uma ferramenta de verificação de qualidade desenvolvida por várias organizações (COPE, DOAJ, STM, etc.) que auxilia investigadores na avaliação da adequação e credibilidade de uma editora, especialmente em ambientes Open Access.

Ferramentas de Workflow e Reprodutibilidade

Devido à complexidade das análises científicas, é essencial documentar os workflows (análises e pipelines) para garantir a sua replicação.

WorkflowHub: É um repositório registado que visa ser o principal catálogo e centro de partilha para todos os workflows científicos computacionais publicamente disponíveis.

Jupyter Notebooks / eNotebooks: São aplicações web de código aberto que permitem aos investigadores criar análises interativas e documentação detalhada da proveniência dos dados, documentando o passo-a-passo exato de como os dados brutos foram processados, limpos e analisados, o que é crucial para a integridade. Um Notebook documenta o passo-a-passo exato de como os dados brutos foram processados, limpos e analisados, contendo código executável e a documentação narrativa.

Binder: É um serviço open source que transforma um repositório de código (por exemplo, um Jupyter Notebook no GitHub ou Zenodo) num ambiente computacional interativo e executável. Permite criar um link que lança o código e os dados para uso direto num browser, instalando todas as dependências necessárias para que o código corra corretamente.

Software e Código-Fonte

Ferramentas como GitHub / GitLab são usadas para código-fonte, oferecendo controlo de versões e permitindo o desenvolvimento transparente e colaborativo. A funcionalidade de integração permite criar uma versão arquivada (release) do código. O Software Heritage é um repositório que visa arquivar e preservar o código-fonte de todo o software publicamente disponível, garantindo a preservação a longo prazo.

Documentação

É essencial incluir documentação detalhada, como README files ou codebooks (livros de código ou dicionários de dados), junto com os dados. Um README file deve descrever a metodologia de recolha, a estrutura de pastas, as convenções de nomes de ficheiros, cabeçalhos das colunas, o significado de valores codificados e unidades de medida.

Ferramentas de Análise (GenAI)

Ferramentas de IA generativa (GenAI), como NotebookLM e ChatPDF, podem ajudar no brainstorming, na revisão de literatura (sumarizar documentos longos, extrair conceitos-chave) e na análise de dados (interpretação de datasets ou documentos como entrevistas anónimas). No entanto, deve ser proibida a introdução de dados críticos (dados pessoais, matérias confidenciais, ideias-chave de investigação) em ferramentas GenAI não aprovadas pela instituição, pois são falíveis e podem gerar erros.

Ferramentas e Serviços da EOSC

Os serviços da EOSC incluem:

- File Sync & Share (armazenamento pessoal e colaborativo na cloud para investigação).
- Large File Transfer (transferência rápida e segura de grandes ficheiros).
- Interactive Notebooks (espaço partilhado para codificação e análise, como JupyterHub).
- Virtual Machines (VMs) (computação em cloud escalável para resultados fiáveis e reprodutíveis).
- Cloud Container Platform (Kubernetes simplificado para investigação, e.g., Docker).
- Bulk Data Transfer (transferências de dados de alto volume entre diferentes Datacenters da EOSC).

O EOSC Resource HUB também oferece um conjunto de recursos, incluindo guias de interoperabilidade.

Publicação Open Access (Open Research Europe – ORE)

A Open Research Europe (ORE) é uma plataforma de publicação científica de Acesso Aberto (Diamante), gratuita e financiada pela Comissão Europeia, destinada a publicações resultantes de projetos financiados pelo Horizonte 2020 e Horizonte Europa. Implementa a revisão por pares aberta. Quando um artigo é submetido, passa por verificações de qualidade básicas, é imediatamente publicado como preprint e enviado para revisão. Esta revisão é totalmente aberta: os revisores são nomeados (não anónimos), os seus pareceres são publicados online com o artigo, e as respostas do autor aos revisores também são publicadas.

Sumário de outputs e exemplos de ficheiros e onde partilhar:

Categoria de Output	Exemplo de Ficheiro (Onde Partilhar)	Porquê é Necessário?
Datasets Finais e Agregados	Ficheiro CSV, Parquet, ou NetCDF (com metadados FAIR).	A base de todas as conclusões.
Software e Algoritmos	Código Python/R (no Zenodo, GitHub ou Repositório Institucional).	A "receita" de como os dados foram processados. Sem o código, os dados são inúteis para validação.
Protocolos e Métodos	Documentação detalhada dos passos de recolha (em PDF, DOCX, etc.).	Garante a Reutilização do método para replicar o estudo em outro contexto.
Workflows e Pipelines	Ficheiro .ipynb (Jupyter Notebook) ou descritores de workflow (CWL).	Documenta a sequência de processamento para garantir a auditoria da análise.
Metadados	Metadados do identificador persistente (PID) da amostra.	Liga a investigação digital à sua fonte física, promovendo a rastreabilidade.

Ferramentas e recurso, por tipo de atividade:

Tipo de Atividade	Ferramentas/Recursos Recomendados	Tipo de Ferramenta
Planeamento e PGD	ARGOS (OpenAIRE) , DMPonline, DS Wizard	Elaboração de PGDs, suporte a maDMPs
	OpenAIRE Costing Tool	Estimativa de custos de RDM
	Fairsharing.org	Referência para standards de metadados disciplinares
Recolha e Segurança	AMNESIA (OpenAIRE)	Anonimização de dados sensíveis
	VeraCrypt / 7-zip	Encriptação e armazenamento seguro
	Jupyter Notebooks / eNotebooks	Documentação de procedimentos e proveniência
Análise e Código	GitHub / GitLab	Versionamento e partilha de código open source
	Binder	Criação de ambientes computacionais reprodutíveis
	Open Research Europe (ORE)	Plataforma de publicação Open Access imediato (HE)
Depósito e Preservação Verificação	Zenodo (OpenAIRE)	Repositório genérico, atribuição automática de DOI
	Re3data	Registo global de repositórios confiáveis (Trusted Repositories)
	OpenAIRE EXPLORE	Descoberta, linking de outputs e reporting
	Think. Check. Submit.	Checklist de qualidade para seleção de editoras/revistas

6.3. Caso de Estudo: Fluxo de Trabalho com Repositório

Em projetos, como um caso de estudo de smart city, com recolha de dados de sensores e criação de software de visualização 3D, o depósito deve incluir não apenas datasets finais e agregados, mas também os outros outputs necessários para a validação e reutilização, como software, algoritmos, protocolos e workflows. Isto cria uma ligação verificável (rastreio de origem) entre a publicação e os processos de trabalho.

Exemplo de fluxo de trabalho com ferramentas, e exemplos de outputs deste caso de estudo:

Fase	Ação essencial	Ferramenta/recurso utilizado	Exemplo de output do projeto
1. Planeamento	PGD Machine-Actionable	Argos (OpenAIRE)	PGD definindo o uso de metadados GeoJSON para as coordenadas dos sensores de trânsito.
2. Recolha e Documentação	Documentação Detalhada e Licenciamento	README File e Licença CC BY 4.0	Ficheiro README.md que descreve o significado de "PM2.5" e a metodologia de calibração do sensor.
3. Desenvolvimento do Código	Controlo de Versões	GitLab	O repositório Git que contém o código-fonte (em Python e JavaScript) da aplicação de visualização 3D.
4. Análise e Workflow	Registo do Processo de Análise	Jupyter Notebooks	02_clean_aggregate.ipynb – O Notebook que converte dados brutos de GPS em fluxos médios de veículos por hora.
5. Execução Imediata	Ambiente Computacional Executável	Binder	Um link de Binder que permite re-executar o Notebook de agregação de dados no navegador.
6. Depósito do Dataset	Atribuição de PID (DOI)	Repositório Institucional da ULisboa	Ficheiro dados_trafego_agregado_2025.csv (com dados anonimizados) com DOI.
7. Depósito do Software	Arquivo e Citação do Código	GitLab para Zenodo Release	Versão 1.0 do software de visualização 3D (smart_city_viz_v1.0.zip) com DOI.
8. Publicação Científica	Publicação Rápida com Revisão Aberta	Open Research Europe (ORE)	O artigo científico sobre o projeto, citando explicitamente os DOIs do Dataset e do Software.
9. Capacitação	Materiais de Formação FAIR-by-Design	Zenodo e CC BY 4.0	O conjunto de slides (com licença CC BY 4.0) utilizado para formar investigadores sobre tratamento de dados de sensores e o software 3D.
10. Preservação	Preservação Perpétua do Código	Software Heritage (SWH)	O SWHID (identificador persistente) do código de visualização 3D, garantindo a sua acessibilidade futura.

Para os materiais didáticos e de formação, a metodologia FAIR-by-Design exige que sejam depositados com DOI (Findable), em Acesso Aberto em formatos abertos (e.g., PDF/A, TXT) (Accessible), usando standards de metadados como o Dublin Core (Interoperable), e publicados com licença Creative Commons CC BY 4.0 (atribuição) (Reusable).

6.4. Obrigações e Boas Práticas

Tema	Práticas Obrigatórias	Boas Práticas (Recomendadas)
Planeamento	Utilizar um PGD (Data Management Plan) como deliverable.	Utilizar ferramentas PGD que suportem a automação (maDMPs), como o ARGOS.
Identificadores	Atribuição de PIDs (DOI para dados, ORCID para autores) a datasets e outputs.	Utilizar PIDs para todos os elementos (software, protocolos) e ROR para instituições.
Partilha/Depósito	Depositar dados (e metadados) num Trusted Repository.	Partilhar outros resultados (código, algoritmos, protocolos, workflows) necessários para a validação.
Metadados	Metadados devem ser FAIR, padronizados e sob licença CC0 ou equivalente.	Usar ferramentas de criação de metadados legíveis por máquinas (e.g., CEDAR).
Segurança, Ética e RGPD	Cumprir o RGPD para dados pessoais, assegurando o Data Protection by Design.	Utilizar AMNESIA para anonimização robusta, retirando os dados do âmbito do RGPD, e encriptação (e.g., VeraCrypt).
Reprodutibilidade	Fornecer informação sobre softwares ou métodos necessários para aceder, validar, e reutilizar os dados.	Partilhar o código-fonte (GitHub/Zenodo) e utilizar plataformas de ambiente computacional reproduzível (e.g., Binder).

7. Ética e Transparência

7.1. Conceitos Chave

A transparência é o pilar da integridade científica, prevenindo má conduta como HARKing (formular hipóteses post-hoc), data dredging (p-hacking, uma prática menos responsável que a avaliação quantitativa impulsionou) e falsificação. O foco ético abrange a proteção de dados pessoais (RGPD, DPIA, consentimento) e a partilha responsável (princípio "tão aberto quanto possível") e o reconhecimento de comunidades vulneráveis (princípios CARE). A reforma da avaliação (CoARA, da qual a ULisboa é signatária desde 2024, e DORA) visa valorizar a qualidade, o impacto e a diversidade de outputs (dados, métodos), afastando-se de métricas meramente quantitativas (JIF, h-index). A Reforma da Avaliação da Investigação (RRA) avança com uma coalizão de signatários (Coalition for Advancing Research Assessment - CoARA) que visa implementar o Acordo para a Reforma da Avaliação da Investigação (ARRA).

Em 2024, surgiu a declaração de Barcelona sobre informação de investigação aberta, que apela ao fim do uso de bases de dados proprietárias e fechadas (como as que alimentam métricas comerciais) para a avaliação de ciência. As instituições são agora incentivadas a utilizar infraestruturas abertas para gerir metadados de publicações e citações, garantindo que a soberania sobre os dados de investigação permanece na mão da comunidade académica e não de editoras comerciais.

O Acesso Aberto influencia diretamente a "velocidade da citação". Estudos indicam que artigos em acesso aberto não só recebem mais citações totais, como são citados muito mais rapidamente após a publicação do que os artigos em acesso fechado. Além disso, a abertura de publicações permite o escrutínio público e mediático, transformando a ciência num recurso para a criação de políticas públicas baseadas em evidência (Evidence-based policy), um impacto que não é captado pelas métricas bibliométricas tradicionais.

A transição para métricas responsáveis exige o uso de Altmetrics (métricas alternativas às quantitativas) para captar o impacto social, político e mediático da investigação que as citações tradicionais ignoram. Deve-se promover o uso de indicadores que reflitam a qualidade intrínseca do trabalho, como a utilização de dados em políticas públicas ou a menção de software em guias clínicos, em detrimento do prestígio da revista onde o artigo foi publicado.

O Pré-registo de estudos é um exemplo de uma prática chave para aumentar a transparência metodológica antes da recolha de dados. A Comissão Europeia foca na CA como uma abordagem para tornar a ciência mais eficiente, produtiva, transparente e robusta, e acelerar o processo de descoberta, melhorar a qualidade da investigação e torná-la mais impactante. Outro exemplo é o Open Peer Review, prática recomendada de CA, que deve ser incentivada para aumentar a transparência, permitindo que as identidades dos revisores sejam conhecidas, os relatórios de revisão sejam publicados e se promova uma discussão pública e contínua após a publicação (post- publication peer review). Isto transforma a avaliação num diálogo colaborativo e não num processo fechado.

A integridade da investigação, a responsabilidade (accountability) e o respeito são princípios fundamentais da ética na investigação.

7.2. Considerações Éticas na Gestão e Partilha de Dados

O princípio central da Ciência Aberta, que orienta a gestão de dados, é: "tão aberto quanto possível, tão fechado quanto necessário". Devem existir justificações válidas para restringir o acesso a dados, como a proteção de direitos humanos, segurança nacional, confidencialidade, direitos de Propriedade Intelectual (PI), conhecimentos secretos, informações pessoais, proteção de espécies ameaçadas ou conhecimentos indígenas. Isto inclui a justificação para não abrir dados quando são comercialmente valiosos se a abertura comprometeria a sua exploração ou dificultaria a proteção de PI. A regra "tão aberto quanto possível, tão fechado quanto necessário" significa que os dados devem ser abertos por omissão, mas fechados quando necessário e com uma justificação válida.

Princípios de Governança e Controlo

Os Princípios CARE (Collective Benefit, Authority to Control, Responsibility, Ethics) são essenciais para a governação de dados, em particular para o conhecimento e dados de comunidades locais e indígenas. Estes princípios garantem que, ao partilhar dados de comunidades vulneráveis ou indígenas, o benefício reverte para a comunidade e que esta tem o direito de controlar o acesso (Authority to Control). O dever fiduciário exige que o Plano de Gestão de Dados (PGD) aborde explicitamente quem terá os direitos de controlar o acesso (Authority to Control) aos dados.

O Dever Fiduciário exige que o Plano de Gestão de Dados (PGD) aborde explicitamente quem terá os direitos de controlar o acesso (Authority to Control) aos dados, especialmente a forma como o conhecimento e dados pessoais de comunidades são utilizados. Em projetos com múltiplos parceiros, este dever deve ser estabelecido no Acordo de Consórcio (AC). O Consentimento Informado deve ser livre, específico, informado e inequívoco, e deve ser obtido antes de iniciar a recolha de dados pessoais. O consentimento deve cobrir a preservação e/ou partilha futura dos dados, quando aplicável.

O Consentimento Informado deve ser livre, específico, informado e inequívoco, e deve ser obtido antes de iniciar a recolha de dados pessoais. O consentimento deve cobrir a preservação e/ou partilha futura dos dados, quando aplicável.

Conformidade Legal (RGPD) e Proteção de Dados

A correta aplicação do Regulamento Geral de Proteção de Dados (RGPD) exige a implementação da proteção de dados desde a conceção e por padrão (*Data Protection by Design and Default*). Os princípios do RGPD incluem a minimização de dados, a transparência e a responsabilidade. O RGPD exige accountability (manter registos) e a proteção de dados deve ser garantida pelo Controlador de Dados.

É obrigatório realizar Avaliações de Impacto sobre a Proteção de Dados (AIPDs ou DPIAs) para atividades de tratamento de dados de alto risco, como o tratamento de dados sensíveis em larga escala. A AIPD é tipicamente exigida quando são cumpridos pelo menos dois critérios de alto risco. O PGD deve prever a anonimização, pseudonimização ou acesso restrito antes do depósito no repositório para garantir a proteção de dados pessoais e sensíveis. As políticas mais robustas exigem que os PGDs abordem explicitamente a proteção de dados pessoais, os interesses comerciais e os direitos de PI. As questões de PI, titularidade e controlo de acesso devem ser cobertas no acordo de consórcio para projetos multi-parceiros. Devem existir diretrizes claras para os Acordos de Partilha de Dados (Data Sharing Agreements - DSAs) mesmo para projetos individuais.

As políticas mais robustas exigem que os PGDs abordem explicitamente a proteção de dados pessoais, os interesses comerciais e os direitos de PI. As questões de PI, titularidade e controlo de acesso devem ser cobertas no acordo de consórcio para projetos multi-parceiros. Devem existir diretrizes claras para os Acordos de Partilha de Dados (Data Sharing Agreements - DSAs) mesmo para projetos individuais.

Integridade da Investigação e Ciência Aberta

A Ciência Aberta (CA) promove a integridade científica quando é baseada nos valores de transparência, escrutínio e crítica. A abertura de dados e métodos serve como garantia contra má conduta, como fabricação, falsificação e plágio. Um PGD ajuda a garantir a consistência, qualidade e documentação dos dados.

A Integridade Científica implica a proibição explícita de má conduta (misconduct), conforme reforçado pelo Embassy of Good Science e outras diretrizes éticas (p.ex: Declaração de Helsínquia da Associação Médica Mundial). A responsabilidade por prevenir e abordar a fraude é partilhada entre investigador e instituição. A integridade deve ser apoiada por uma Cultura de Investigação segura, justa e de suporte.

Devem existir planos para cenários excecionais; por exemplo, o projeto PREPARED criou um código para garantir que a investigação em tempos de crise ou pandemia é conduzida de forma ética e eficiente, sem comprometer os princípios de integridade.

A consulta a comités de ética é obrigatória quando o projeto envolve pessoas ou animais, ou incide sobre o ambiente. A integridade é reforçada quando a abertura abrange todos os resultados da investigação, incluindo software, algoritmos, protocolos e workflows.

O uso de Inteligência Artificial (IA) levanta novas questões éticas e de transparência, sendo necessário ter consciência do viés (bias) algorítmico e da opacidade dos métodos (opacity) dos modelos de IA. O uso de IA deve incluir a preservação de registos claros (logs) para fins de auditoria e rastreabilidade (accountability), o que é uma responsabilidade de gestão.

Eixos Éticos e Implicações na Gestão de Dados (RDM)

1. Pessoas e Comunidades:

- Dados Pessoais e Sensíveis (RGPD): O PGD deve prever a anonimização, pseudonimização ou acesso restrito antes do depósito. O consentimento deve cobrir a futura partilha e reutilização dos dados.
- Ética Social e Dados de Comunidade (CARE): Adotar os Princípios CARE, garantindo que a comunidade vulnerável ou indígena tem o direito de controlar o acesso (Authority to Control).
- Evitar Enviesamentos e a Discriminação: Documentar e mitigar os enviesamentos

inerentes à recolha de dados, pois a falta de representatividade dos dados pode levar a conclusões injustas.

2. Processo e Integridade:

- Transparência e Reprodutibilidade: Partilhar os dados de suporte (FAIR Data) e o código/workflow (FAIR Software) usado, pois a ocultação de dados compromete a integridade científica.
- Prevenção de Má Conduta (Misconduct): O Pré-registo de estudos e a abertura dos protocolos (em plataformas como o OSF) previnem práticas questionáveis como p-hacking ou manipulação de dados após a análise.
- Uso Ético de IA: Proibir introdução de dados confidenciais em modelos de IA não aprovados. Exigir declaração de uso de GenAI nos relatórios e manter logs do uso de IA para rastreabilidade, assegurando a responsabilidade humana final.

3. Reconhecimento e Avaliação:

- Atribuição Justa de Crédito: Atribuir um DOI ao Dataset e citá-lo formalmente no artigo. Garantir que os Data Stewards e Data Managers são reconhecidos no PGD.
- Reconhecimento do Software e Workflow: Dar crédito ao software e aos workflows computacionais (ex: usando DOI do Zenodo ou registo no WorkflowHub). O software é um output de investigação, não apenas uma ferramenta.
- Avaliação (CoARA/DORA): A avaliação deve focar-se na qualidade, impacto e práticas abertas (incluindo a partilha de dados e software), e não apenas em métricas quantitativas de publicação. Práticas menos responsáveis podem ser incentivadas por fatores de impacto meramente quantitativos, que podem levar à crise da replicação.

7.3. Integridade Científica e Protocolos Éticos

O pré-registo (pre-registration) de estudos é uma prática recomendada para aumentar a transparência, publicando o plano de estudo antes de a recolha de dados ser iniciada. O

pré-registo (Pre-registration) permite a publicação do plano de estudo, incluindo hipóteses e análise estatística, em plataformas públicas e imutáveis antes do início da coleta.

O pré-registo de relatórios registados (registered reports) envolve a submissão do desenho da investigação, introdução, métodos e plano de análise para peer review em uma revista científica antes da recolha de dados. Se o plano for aceite, a revista compromete-se a publicar o artigo, independentemente do resultado.

A Comissão Europeia recomenda o pré-registo, relatórios registados (registered reports) e depósito de dados em repositórios partilhados. A retenção de dados para fins de verificação é importante, mesmo para dados não publicados.

A má conduta na investigação (misconduct) também inclui a criação de citações falsas geradas por IA.

Tipos de Má Conduta

A transparência visa prevenir má conduta, que inclui:

- Data dredging (p-hacking): Analisar dados para encontrar resultados estatisticamente significativos sem uma hipótese predefinida, o que pode levar a conclusões enganadoras e falsos positivos.
- HARKing (Hypothesizing After the Results are Known): Formular hipóteses depois de conhecer os resultados, apresentando-as como se fossem originais, a priori. Isto pode distorcer o processo de investigação e levar a um registo científico tendencioso.
- Falsificação: Inventar dados, resultados ou observações, ou manipular resultados (p.ex., remover outliers sem justificação).
- Plágio: Copiar textos, ideias, métodos ou resultados sem atribuição adequada.
- Autoria Indevida: Incluir autores que não contribuíram (gift authorship) ou excluir autores que contribuíram (ghost authorship).
- Manipulação de Imagens: Ajustes excessivos ou enganosos em figuras (p.ex., manipular a resolução de imagens microscópicas).
- Destruição de Dados: Apagar dados deliberadamente para evitar replicação ou

auditoria.

- Seleção de Resultados (Cherry-picking): Publicar apenas resultados positivos e omitir negativos ou neutros; selecionar apenas evidências que suportam a conclusão desejada.
- Má Descrição de Métodos/Mau Registo de Dados: Falta de detalhes ou de documentação adequada que impeça a replicação; falta de metadados.
- Não Reportar Limitações: Esconder fragilidades ou enviesamentos do estudo.
- Confundir Correlação com Causalidade: Apresentar relações correlacionais como determinísticas.
- Utilizar Análises Inadequadas: Aplicar testes estatísticos errados, violando pressupostos.
- Consentimento Inadequado ou Enganoso: Participantes não entendem os riscos, direitos ou forma de uso dos dados.
- Falta de Proteção de Dados Sensíveis: Risco de reidentificação, má anonimização, ou partilha indevida de dados brutos.
- Exploração de Comunidades Vulneráveis: Investigações que beneficiam investigadores, mas não a comunidade estudada.
- Coerção ou Incentivos Excessivos: Fazer os participantes sentirem que devem participar ou incentivar materialmente, criando distorções.
- Ocultar Conflitos de Interesse: Existência de relações financeiras ou institucionais não declaradas.
- Publicação Redundante (Self-plagiarism): Publicar o mesmo estudo em vários locais diferentes.
- Salami Slicing: Dividir artificialmente um estudo em múltiplas publicações para inflacionar métricas.
- Manipulação no Peer Review: Colocar revisores falsos, enviar revisões fraudulentas, ou sabotar colegas.

- Citações Estratégicas: Citar artigos sem relevância apenas para agradar a editores ou elevar métricas.

Enquadramento de Iniciativas e Diretrizes Éticas

Diversas iniciativas visam reforçar a integridade científica no contexto da CA:

Fonte / Iniciativa	Foco Principal	Princípios Éticos Chave	Aplicação ao Contexto do Estudo (dados sensíveis, comunidades vulneráveis)
COARA – Coalition for Advancing Research Assessment	Reformar práticas de avaliação da investigação	- Avaliação baseada em qualidade e impacto, não apenas métricas - Transparência nos critérios - Reconher de contributos diversos	Incentiva avaliação ética de trabalhos com dados sensíveis, valoriza abordagens responsáveis e não métricas quantitativas derivadas de datasets restritos.
DORA – Declaration on Research Assessment	Redução do uso indevido de métricas (ex.: Journal Impact Factor)	- Avaliar a investigação pelos seus méritos - Evitar métricas simplistas - Reconhecer práticas abertas e contributos diversos	Reforça que restrições éticas nos dados não prejudicam a avaliação da investigação; valoriza práticas como documentação, qualidade metodológica e ética.
European Code of Conduct for Research Integrity (ALLEA)	Padrões fundamentais para integridade científica na UE	- Fiabilidade - Honestidade - Respeito - Responsabilidade - Boas práticas de gestão de dados e consentimento	Impõe rigor no tratamento de dados pessoais; exige anonimização robusta, consentimento informado, segurança, e comunicação responsável dos resultados.
ROSiE – Responsible Open Science Guidelines	Harmonizar ciência aberta com ética e segurança	- Minimizar riscos para participantes - Princípio "as open as possible, as closed as necessary" - Gestão responsável de dados sensíveis - Avaliação de riscos de reidentificação	Orienta que dados de estudos sensíveis/pessoais não sejam abertos, apenas metadados; recomenda ferramentas como Amnesia e revisão ética contínua.
Embassy of Good Science (incl. VIRT2UE, CHANGER, PREPARED Code)	Princípios éticos práticos para investigadores	- Virtudes científicas (integridade, justiça, coragem, prudência) - Tomada de decisões éticas em contextos complexos - Prevenção de má conduta científica - Condutas responsáveis em ambientes vulneráveis	Fornece diretrizes para proteger participantes vulneráveis, promover empatia, justificar restrições no acesso a dados e evitar danos sociais.
COPE – Committee on Publication Ethics (Guidelines for Good Publication Practice)	Boas práticas na publicação científica	- Integridade na autoria - Transparência em métodos e dados - Declarações de conflito de interesse - Gestão ética de dados ilícitos ou sensíveis - DAS adequado (Data Availability Statement)	Suporta a opção de disponibilizar apenas metadados; exige transparência sobre restrições; orienta como escrever a DAS e como proteger identidades em publicações.

A síntese dos requisitos comuns de protocolos éticos inclui a proteção de participantes vulneráveis, a gestão responsável de dados sensíveis (anonimização, acesso restrito, metadados FAIR), a transparência metodológica, a ética na autoria e publicação (COPE) e a aplicação do princípio da ciência aberta responsável ("tão aberto quanto possível, tão fechado quanto necessário").

7.4. Casos de Estudo: Dilemas Éticos na Partilha

Cenário: Investigação em Comunidades Vulneráveis

Num cenário de investigação social sobre violência doméstica em comunidades vulneráveis, são recolhidos dados qualitativos e quantitativos, como entrevistas gravadas, dados socioeconómicos e geolocalização aproximada das residências. A anonimização robusta deve ser aplicada em dados pessoais como os de mapeamento participativo, onde a geolocalização torna a reidentificação possível. No caso de dados de fala (speech data), a privacidade pode ser violada se informações secundárias (como condições de

saúde, emoções ou ambiente) forem analisadas para fins não consentidos, mesmo que o conteúdo da fala seja público. A tecnologia avança e pode levar a futuras ameaças de reidentificação.

A partilha de dados brutos neste cenário pode violar princípios éticos e o RGPD. O dataset simulado para fins educativos, embora fictício e artificialmente reduzido em sensibilidade, requer a remoção total de nomes, datas, moradas, e quaisquer detalhes de incidentes, mantendo apenas temas amplos e descrições genéricas. A utilização recomendada para este tipo de dataset inclui formação em ética da investigação e demonstrações de curadoria, documentação ou anonimização, sendo proibida qualquer tentativa de inferir ou associar a casos reais.

Soluções Éticas para Partilha Condicional

1. Anonimização Robusta: Se a abertura total não for possível devido à sensibilidade dos dados, a anonimização robusta deve ser aplicada para tornar a reidentificação impossível, retirando os dados do âmbito do RGPD e permitindo a partilha responsável. A ferramenta AMNESIA do OpenAIRE pode ser usada para anonimização, e pode ser usada on-premise para maior segurança, sem enviar dados pessoais pela Internet. As funcionalidades do AMNESIA incluem a generalização (substituir valores específicos por agregados usando hierarquias), supressão de registos e pseudonimização/masking.

2. Acesso Controlado: Para dados sensíveis que não podem ser anonimizados (ex: vídeos das entrevistas), a solução ética é o acesso controlado (dados fechados). Isto é gerido através de Acordos de Partilha de Dados (DSAs) e exige um procedimento de acesso (managed access procedure) em vigor no repositório confiável. A solução de Acesso Controlado deve ser formalizada através de Acordos de Partilha de Dados (Data Use Agreements - DUAs), que regulam o acesso restrito.

Transparência com Dados Fechados (DAS)

Mesmo para dados fechados, existe o dever de transparência: a Data Availability Statement (DAS) deve descrever explicitamente as condições e a localização (PID) do dataset, mesmo que este seja restrito e que apenas os Metadados estejam disponíveis. O dataset restrito pode disponibilizar apenas metadados públicos e dados altamente anonimizados mediante pedido. O acesso será concedido apenas após avaliação ética e assinatura de acordo de utilização restrita.

Casos: Ética em Publicação e Fontes Públicas

Reconhecimento de Serviços de Redação: O custo de um serviço profissional de escrita não anula a obrigação de transparência e honestidade científica. O reconhecimento deve ser dado nos agradecimentos para garantir que o leitor sabe quem esteve envolvido no processo de redação e para evitar suspeitas de ghostwriting.

Peer Review e Confidencialidade: O peer review baseia-se na confiança. Os autores devem ter a certeza de que o seu trabalho, confidencial e não publicado, será utilizado apenas para a avaliação de mérito científico. A cópia de texto de manuscritos em revisão para uso pessoal ou de terceiros é uma má prática.

Plágio em Longo Prazo: Casos de plágio podem ser descobertos e levados a retratação anos após a publicação.

Dados Pessoais Online: Dados pessoais disponíveis em redes sociais, websites, fóruns ou bases de dados abertas continuam a estar protegidos pelo RGPD. O facto de algo ser público (estar publicado) não significa que possa ser reutilizado livremente para investigação. O RGPD exige que os investigadores avaliem se os indivíduos tinham uma expectativa razoável de privacidade.

7.5. Obrigações e Boas Práticas

Tema	Práticas Obrigatórias	Boas Práticas Opcionais (Recomendadas)
Conformidade Legal (RGPD)	Cumprir o RGPD e os princípios de minimização de dados. Obter Consentimento Informado para a recolha, preservação e partilha de dados pessoais.	Realizar Avaliações de Impacto sobre a Proteção de Dados (AIPDs) para refletir sobre os riscos e as salvaguardas.
Abertura Condicional	Partilhar dados sob o princípio "Tão aberto quanto possível, tão fechado quanto necessário". Justificar a não partilha no PGD.	Aplicar a Anonimização robusta (e.g., usando AMNESIA) para que os dados saiam do âmbito do RGPD e possam ser abertos.
Integridade Científica	Submeter a metodologia a Revisão Ética (Comité de Ética) em todos os casos aplicáveis.	Pré-registo de estudos (em plataformas abertas) para aumentar a transparência e prevenir má conduta. Partilhar software, algoritmos, protocolos e workflows.
Governação e Controlo	Garantir que o PGD aborda explicitamente a conformidade legal, PI e propriedade dos dados.	Adotar os Princípios CARE (Collective Benefit, Authority to Control) para dados de comunidades vulneráveis. Formalizar Data Sharing Agreements (DSAs) para dados confidenciais.
IA e Autores	O investigador é responsável pela integridade dos outputs. Exigir declaração obrigatória do uso de IA.	Proibir o input de dados críticos (pessoais, confidenciais) em ferramentas GenAI não aprovadas pela instituição. Manter registos claros (logs) do uso de IA para transparência e rastreabilidade.
Partilha Restrita	Garantir que os Metadados sejam sempre abertos (sob CC0) e FAIR, mesmo que o dataset seja fechado.	Formalizar Acordos de Partilha de Dados (DSAs) para gerir o acesso controlado e seguro a dados sensíveis e/ou pseudonimizados.

8. Reprodutibilidade

8.1. Conceitos Chave

A reprodutibilidade é a capacidade de obter os mesmos resultados usando os mesmos dados e código (verificação) e é essencial na gestão e partilha de dados para garantir a validade e replicabilidade dos resultados científicos. Exige a partilha abrangente de todos os outputs (dados, código, protocolos, workflows e ambientes computacionais). A documentação da proveniência (histórico detalhado das transformações dos dados) é fundamental e é um requisito dos PGDs.

O problema das configurações de software (dependencies) é resolvido pela tecnologia de contentores e ambientes virtuais (Jupyter/Binder), isto para garantir que o software corre exatamente da mesma forma daqui a 10 anos. Esta técnica encapsula o código, as bibliotecas e o sistema operativo num único ficheiro, eliminando o erro clássico de depender de uma máquina com configurações específicas, e permitindo que qualquer investigador possa replicar a análise computacional com uma instalação simplificada.

A reprodutibilidade é a base para iniciativas como o registered reports (relatórios registados) ou pre-registration (pré-registo) de estudos. A interoperabilidade legal (CODATA-RDA) garante que a reutilização de dados de diferentes fontes é compatível legalmente (licenças). A reprodutibilidade aumenta a confiança do público na ciência e fortalece a integridade científica, além de aumentar a eficiência e o impacto da investigação.

A reprodutibilidade é uma prática obrigatória, p.ex. no Horizonte Europa, e exige a garantia de que a comunidade científica possa obter os mesmos resultados usando os mesmos métodos, ferramentas e dados. Os resultados que os beneficiários devem disponibilizar para validação e reutilização podem incluir software, algoritmos, protocolos, modelos, workflows e electronic notebooks (cadernos eletrónicos). Para garantir a reprodutibilidade, é necessário fornecer acesso aos dados/resultados para validação e reutilização, mas isto não basta: é crucial documentar e partilhar também o fluxo de trabalho de dados (data workflow), incluindo os dados brutos (raw data), os scripts de limpeza e os parâmetros de software utilizados. A reprodutibilidade técnica exige que um investigador independente consiga obter os mesmos resultados utilizando os mesmos materiais e métodos. As práticas de Early Sharing (Pré-registo e Preprints) e o uso de Planos de Gestão de Dados (PGDs) interoperáveis são recomendadas para

aumentar a transparência e a reprodutibilidade.

A nova fronteira é nos dados prontos para IA (*AI-Ready Data*). Isto implica que, além dos princípios FAIR, os dados devem ser acompanhados de "Model Cards" ou "Datasheets for Datasets", que descrevem os enviesamentos (biases) presentes nos dados, os métodos de treino utilizados e as limitações éticas da sua aplicação em algoritmos de inteligência artificial, evitando a propagação de preconceitos automatizados.

A reprodutibilidade é um dos aspetos centrais do Responsible Conduct of Research (RCR).

8.2. O Papel dos Dados na Reprodutibilidade Científica

Reprodutibilidade: Significa obter resultados idênticos ou semelhantes utilizando os mesmos dados e o mesmo código/métodos de análise (isto é, uma verificação dos resultados científicos originais).

Replicabilidade: Significa obter resultados semelhantes utilizando o mesmo método, mas novos dados (isto é, uma validação através da replicação do mesmo método, alargando a validade dos resultados científicos para um novo contexto).

A reprodutibilidade exige a garantia da integridade e proveniência dos dados. A Proveniência, em particular, requer metadados detalhados sobre as transformações, versões e o registo histórico da origem dos dados, incluindo o tempo, as operações realizadas, os dados de entrada, os parâmetros usados e quem tratou os dados. O repositório deve suportar o versionamento dos dados e o rastreamento da sua proveniência (provenance) para que a rastreabilidade seja transparente.

Princípios FAIR e Reprodutibilidade

Os Princípios FAIR (Findable, Accessible, Interoperable, Reusable) estão diretamente ligados à reprodutibilidade:

Princípio FAIR	Foco na Reprodutibilidade	Elementos relevantes para a GDI
F (Findable)	Acesso rápido a todos os outputs	PIDs (DOIs) para dados, código, workflows. Metadados ricos e sob licença CC0.
A (Accessible)	Garantir que os softwares necessários estejam disponíveis.	Especificar softwares ou métodos necessários para reusar, incluir links (PID) na Data Availability Statement (DAS).
I (Interoperable)	Garantir que dados e código possam ser lidos por diferentes sistemas (máquinas).	Uso de (CSV, PDF/A). Uso de metadados standard e/ou específicos do domínio.
R (Reusable)	Garantir a Proveniência, a Autoria e as Condições de Reutilização.	Proveniência documentada (versionamento do código) e PIDs de agentes/ferramentas. Licença clara (CC BY ou CC0). Documentação detalhada (README files, codebooks). Critérios CR 1a e CR 1b do Science Europe.

Avaliação da Proveniência e Qualidade dos Dados (Guia Science Europe)

O Guia da Science Europe para Revisores de PGDs (Core Requirements for DMPs) fornece critérios detalhados para avaliar a Proveniência e Qualidade dos Dados, que são essenciais para a reprodutibilidade. Esta avaliação de DMPs centra-se na análise da cobertura de todos os aspetos necessários:

Requisito Core (CR)	Foco	Documentação Exigida no PGD
CR1. Descrição dos dados	Recolha ou reutilização de dados existentes.	O PGD deve detalhar como serão recolhidos ou produzidos novos dados e/ou como serão reutilizados dados existentes. Deve especificar que dados (tipos, formatos e volumes) serão recolhidos ou produzidos.
CR2. Documentação e qualidade dos dados	Qualidade e documentação.	O PGD deve indicar que metadados e documentação vão acompanhar os dados (por exemplo, a metodologia de recolha e a forma de organização dos dados). Devem ser definidas as medidas de controlo de qualidade dos dados a serem utilizadas.

Critérios de Revisão Adicionais (Proveniência e Qualidade):

- Consistência e Completude: Os dados são consistentes e completos? É necessário justificar as omissões (missing data), como, por exemplo, por que razão 15% das amostras de trânsito foram descartadas devido a falha de conectividade do sensor. O DMP deve listar todos os passos de transformação sofridos pelos dados, desde a recolha bruta até ao output final.

- Validação da Precisão: Devem estar documentadas as margens de erro e os níveis de incerteza inerentes à recolha (ex: ruído de fundo que afetou os sensores).

- Rastreabilidade do Processo: É necessário verificar se a proveniência dos dados está completa. O PGD deve listar todos os passos de transformação sofridos pelos dados, desde a recolha bruta até ao output final. Isto inclui as versões específicas do software e os IDs (ORCIDs) dos agentes e/ou ferramentas (DOIs) responsáveis por essas ações.

- Clareza da Documentação: A documentação deve ser suficientemente clara para que os dados sejam entendidos sem ser necessário contactar os autores. Por exemplo, um Codebook deve estar completo e a descrição das variáveis deve ser inequívoca.

- Licenciamento e Acesso: É necessário usar licenças claras e, para dados sensíveis, validar se o Acordo de Partilha de Dados (DSA) proposto é robusto para equilibrar a ética (proteção de dados pessoais) com o acesso científico (reutilização).

- Acessibilidade e Formato: Os revisores devem confirmar se os dados estão num formato padronizado e legível por máquina (Ex: CSV em vez de XLS) e se os PIDs para os dados e metadados estão corretamente atribuídos no repositório. O repositório deve suportar o versionamento dos dados e o rastreamento da sua proveniência (provenance) para que a rastreabilidade seja transparente.

Exemplos de Documentação de Qualidade (Projeto Smart City)

O PGD deve documentar a qualidade dos dados seguindo eixos específicos, como ilustrado no exemplo de um projeto Smart City:

Eixo de Qualidade	O que o PGD Deve Documentar	Exemplo (Projeto Smart City)
Precisão e Exatidão	O grau em que os dados representam o fenómeno real e as margens de erro (uncertainty).	Documentação da tolerância e intervalo de calibração dos sensores de Qualidade do Ar (p.ex., erro de ±5% na leitura de PM10).
Consistência	A ausência de contradições nos dados ao longo do tempo ou entre diferentes fontes.	Garantir que os códigos de Georreferenciação (p.ex., das câmaras, sensores) são uniformes entre os diferentes dias de recolha.
Completude (*Completeness*)	O grau em que faltam dados, e a justificação para as omissões (missing data).	A justificação de por que razão 15% das amostras de trânsito foram descartadas (p.ex., por falha de conectividade do sensor, dados não são fiáveis).
Confiabilidade (*Reliability*)	O grau em que o processo de recolha (método, instrumentos) produziria os mesmos resultados se repetido.	Detalhar que o algoritmo de filtragem de dados de trânsito foi executado duas vezes com os mesmos parâmetros, produzindo o mesmo output.

8.3. Disponibilização de Dados e Resultados para Validação

O princípio da partilha abrangente exige que se forneça informação sobre resultados e qualquer outra ferramenta ou instrumento necessário para validar ou reutilizar os dados de investigação. Isto inclui: software e código-fonte, algoritmos e modelos, protocolos e processos de trabalho (Workflows), e registos de equipamentos de laboratório (logs). Financiadores como o Horizonte Europa exigem que os beneficiários forneçam acesso (digital ou físico) aos outputs de investigação, ferramentas e instrumentos necessários para validar as conclusões das publicações científicas e reutilizar os dados. Isto deve ser feito salvaguardando-se os interesses legítimos ou restrições.

Ferramentas de Apoio à Reprodutibilidade

RDMKit (ELIXIR): Um toolkit integrado que fornece informação contextualizada e guias para cada etapa do ciclo de vida dos dados (planeamento, recolha, processamento, análise, preservação, partilha). Também oferece um catálogo de ferramentas, padrões e guias relevantes, e detalhes sobre esquemas de metadados adequados para áreas científicas específicas (ex: BioSchemas ou padrões para dados de sensores).

Guias de Data Management da Cornell University: Oferecem templates para a criação de ficheiros README eficazes com convenções disciplinares. Um README para dados deve

incluir: 1) o contexto da investigação (hipótese e objetivo); 2) a estrutura dos dados (organização e relacionamento entre ficheiros); 3) a chave de dados (Codebook); e 4) as condições de acesso, a licença e restrições de uso. O README garante que os datasets podem ser compreendidos sem ser necessário consultar o artigo científico. O README file deve ainda descrever a metodologia de recolha, a estrutura de pastas e as convenções de nomes de ficheiros.

Transparência e Processo de Descoberta: Para garantir a transparência no processo de descoberta, ferramentas como o The Lens permitem a exportação em massa (bulk export) de até 1000 resultados de pesquisa e metadados associados, o que é essencial para revisões sistemáticas e estudos de meta-análise que pretendam documentar o fluxo de seleção de literatura de forma auditável.

Exemplo de README (Projeto Smart City)

Este exemplo de ficheiro README ilustra a documentação necessária para a reprodutibilidade, incluindo:

- Descrição: Dados agregados sobre a qualidade do ar (PM10) e fluxo de trânsito veicular em Lisboa.
- Identificador: PID (DOI) do Repositório ULisboa deve ser inserido.
- Metodologia e Proveniência: Deve ser detalhada a fonte dos dados (ex: sensores de baixo custo calibrados) e os passos de processamento. No exemplo, os dados brutos de GPS foram agregados em blocos de 1 hora e médias de 50m de área para garantir o anonimato. Deve ser registada a Filtragem (remoção de dados anómalos de PM10) e a Conversão (variáveis de tempo convertidas para GMT+0:00).
- Estrutura de Ficheiros: Deve listar o ficheiro CSV principal e a documentação associada, como o Codebook_SmartCity_v1.pdf (dicionário de dados) e o script análise_workflow_v1.0.ipynb (o Jupyter Notebook usado para gerar o dataset a partir dos dados brutos).
- Licenciamento: A licença deve ser clara (ex: CC BY 4.0), permitindo a reutilização, partilha e adaptação, incluindo para fins comerciais, desde que o crédito seja atribuído (citando o DOI).

- Dicionário de Dados (Codebook): Deve descrever cada coluna, incluindo o nome, a descrição, a unidade de medida (ex: µg/m³ ou Veículos/hora) e o formato de dados.

Interoperabilidade Legal (CODATA-RDA)

A interoperabilidade legal é a capacidade de combinar e reutilizar dados de diferentes fontes sem violar leis de Propriedade Intelectual (PI), privacidade (RGPD) ou termos de licenciamento. O problema dos silos legais ocorre quando datasets estão legalmente bloqueados (licenças incompatíveis ou conflitos entre leis de diferentes países, p. ex: RGPD vs. EUA).

O CODATA (Committee on Data) e a RDA (Research Data Alliance) indicam diretrizes para garantir que os dados são legalmente reutilizáveis:

1. Clareza (Clarity): As condições legais (licença, restrições) devem ser explícitas, claras e fáceis de entender por um humano. É recomendável evitar linguagem jurídica complexa no PGD.

2. Legibilidade por Máquina (Machine-Readability): As licenças devem ser expressas num formato que o software possa ler e interpretar automaticamente (ex: tags de licença do Creative Commons).

3. Compatibilidade (Compatibility): O PGD deve justificar a escolha de licenças que sejam compatíveis com outros datasets da mesma área (preferir CC BY 4.0 ou CC0 por serem as mais interoperáveis).

4. Harmonização Transfronteiriça: As licenças devem ser agnósticas da jurisdição (globais, como as CC) e deve haver clareza sobre como lidar com conflitos entre leis de diferentes países.

Documentação Abrangente e Transparência

A The Turing Way é uma iniciativa que cria guias comunitários para a ciência aberta e reprodutível, focando na organização de dados (estrutura de diretórios e convenções de nomeação de ficheiros) para garantir a legibilidade. O guia enfatiza o uso de práticas de codificação limpa e controlo de versão, e orientações exaustivas para documentação.

O pré-registo (Pre-registration) de estudos, realizado em plataformas como o Center of Open Science (COS) prereg, é uma prática que aumenta a transparência. Consiste em

especificar a metodologia e o plano de análise de dados antes de a recolha de dados se iniciar, registando-o numa plataforma pública e imutável. O pré-registo é o ato de publicar o plano de estudo, incluindo hipóteses, desenho da investigação e plano de análise estatística, em plataformas públicas e imutáveis antes do início da coleta dos dados ou do exame dos dados.

Tipos de Pré-registo: O pré-registo completo detalha a hipótese, o desenho experimental, o plano de recolha e todas as etapas de análise estatística. Um pré-registo de relatórios registados (registered reports) é um formato mais avançado em que o plano de estudo é submetido a uma revista para peer review antes da recolha de dados, e a revista compromete-se a publicar o artigo independentemente dos resultados (combate o viés de publicação).

O pré-registo melhora a Replicabilidade, dando acesso exato à metodologia, e a Reutilização, pois o conhecimento a priori sobre o plano de análise aumenta a confiança na proveniência e integridade dos resultados.

Exemplo de modelo de pré-registo (Projeto Smart City)

Sumário: O exemplo de um modelo de pré-registo de estudo observacional (Correlação Trânsito/PM10 no Lumiar) detalha a descrição do estudo, as Hipóteses (Primária, Secundária e Nula), o Desenho Experimental (observacional, correlacional, de séries temporais), o Plano de Amostragem (fonte, regra de paragem, e o momento em que a análise confirmatória começará), as Variáveis (Independente, Dependente e Covariáveis) e o Plano de Análise Estatística (Software, Preparação de Dados, Teste da H1 com GLM, Teste da H2 com Correlação Cruzada, e Nível de Significância fixado em $p < 0.05$).

Exemplo completo:

Pré-Registo (OSF Preregistration Template)

1. Título do Estudo: Pré-registo: A Correlação entre o Fluxo de trânsito Veicular e as Concentrações de PM10 no Bairro do Lumiar, Lisboa.

2. Autores e Afiliações:

• *Dr. XXXXX (Investigador Principal, nome da instituição)*

• *Dra. YYYYY (Candidata a Doutoramento, nome da instituição)*

3. Descrição do Estudo (Summary): Este estudo observacional utilizará dados de sensores (IoT) para testar a relação entre a densidade de trânsito de veículos e a poluição do ar (PM10) numa área urbana de Lisboa (Lumiar). O objetivo é quantificar esta relação e identificar eventuais lags temporais entre os picos de trânsito e os picos de poluição.

4. Hipóteses (Confirmatórias):

- *H1 (Primária): Um aumento no fluxo médio de veículos por hora está positivamente correlacionado com um aumento nas concentrações médias horárias de PM10.*
- *H2 (Secundária):O pico de poluição (PM10) ocorrerá com um atraso (lag) de 1 a 2 horas após o pico de trânsito matinal (08:00-10:00).*
- *H0 (Nula): Não existe correlação estatisticamente significativa (p > 0.05) entre o fluxo de veículos e as concentrações de PM10.*

5. Desenho Experimental (Design Plan):

Estudo observacional, correlacional, de séries temporais. Não há manipulação de variáveis; apenas recolha passiva de dados.

6. Plano de Amostragem (Sampling Plan):

Fonte de Dados: 5 sensores de qualidade do ar (Modelo [XYZ]) e 5 sensores de trânsito(indutivos) instalados em locais fixos no Lumiar.

Regra de Paragem (Stopping Rule): A recolha de dados decorrerá continuamente de 1 de janeiro de 2026 a 30 de junho de 2026 (6 meses).

A análise confirmatória só começará após 1 de julho de 2026.

7. Variáveis:

Variável Independente (Preditora): traffic_count (Contagem média horária de veículos, tipo: Inteiro).

Variável Dependente (Resultado): pm10_avg (Concentração média horária de PM10, em µg/m³,

tipo: Float).

Covariáveis (Controlo): hora_do_dia (Hour), dia_da_semana (Binary, Fim-de-semana/Dia útil), velocidade_vento_avg (Float), precipitacao_mm (Float).

8. Plano de Análise Estatística:

Software: A análise será conduzida em R (versão 4.3 ou superior) e RStudio, utilizando os pacotes dplyr para manipulação e lme4 para modelação.

Preparação de Dados: Os dados serão agregados em médias horárias. Outliers (definidos como leituras de PM10 > 3 desvios padrão da média diária) serão removidos da análise confirmatória.

Teste da H1: Será construído um Modelo Linear Generalizado (GLM)para testar o efeito de traffic_count em pm10_avg, controlando pelas covariáveis listadas (vento, dia da semana).

Teste da H2: Será utilizada uma Análise de Correlação Cruzada (Cross-Correlation) para identificar o lag temporal entre as duas séries temporais.

Nível de Significância: O nível de significância (alfa) para todas as análises confirmatórias será fixado em $p < 0.05$.

Reconhecimento e Contribuições (CRediT Taxonomy)

A CRediT Taxonomy do NISO é um sistema normalizado de 14 papéis que define as contribuições individuais para um projeto. É obrigatória em muitas revistas e promove a transparência e a responsabilidade (accountability) na autoria, garantindo que todas as contribuições, incluindo as relacionadas com Dados e Software, sejam reconhecidas. A CRediT Taxonomy é um dos indicadores de integridade científica que deve ser cumprido, e a sua utilização reforça a transparência na autoria.

A Curadoria de Dados (Data Curation)

As ações para anotar, limpar e manter os dados utilizáveis (FAIR) e o Software (aplicação, desenvolvimento ou codificação) são papéis essenciais que devem ser reconhecidos em cada fase da GDI:

Categoria	Papéis Essenciais	Foco em RDM/Reprodutibilidade
Conceção (Conceptualização)	Conceptualização, Metodologia, Aquisição de Financiamento.	O quê e porquê da investigação.
Execução & Análise	Investigação, Curadoria de Dados (Data Curation), Análise Formal, Software.	Curadoria de Dados: Ações para anotar, limpar e manter os dados utilizáveis (FAIR). Software: Aplicação, desenvolvimento ou codificação.
Redação & Revisão	Redação do Rascunho, Redação – Revisão e Edição.	Revisão crítica e edição da versão final do manuscrito.
Gestão & Supervisão	Supervisão, Administração do Projeto, Visualização.	Supervisão: Responsabilidade de gestão e orientação sobre o projeto.

O Desafio do Ambiente Computacional

A simples partilha de scripts em texto é insuficiente para a reprodutibilidade devido a problemas de dependência (bibliotecas, packages). A solução é encapsular o ambiente computacional através da tecnologia de contentores.

Jupyter Notebooks: São aplicações web de código aberto que combinam código, equações e visualizações num único documento, permitindo aos investigadores criar análises interativas e reprodutíveis, documentando o passo-a-passo do processamento e análise dos dados.

Binder: É uma ferramenta que cria um link partilhável que lança o código (ex: um Jupyter Notebook num repositório Git) num ambiente interativo diretamente num Browser. Isto permite a validação e reutilização imediata dos métodos de investigação aplicados, facilitando a replicação/reutilização.

8.4. Casos de Estudo: Replicação e Reprodução de Resultados

Exemplos de Sucesso:

- Física de Partículas (CERN/LHC): Exige um nível de evidência muito rigoroso, de 5σ (1

em 3,5 milhões de probabilidade de erro), com múltiplos detetores independentes (ATLAS, CMS) e revisões exaustivas das publicações.

- Astronomia: Grandes projetos abertos (ex: dados do Hubble, SDSS) publicam dados e códigos que permitem reanálises públicas, e as descobertas são confirmadas por múltiplos grupos independentes.
- Ciência Computacional/Biologia: Domínios que trabalham com grandes volumes de dados adotaram o modelo de investigação reproduzível: código aberto, dados públicos e workflows documentados.

Exemplo de Falta de Replicabilidade:

Um estudo da Open Science Collaboration analisou a reprodutibilidade de 100 estudos publicados em revistas de alto nível em Psicologia. Apenas 36% das replicações produziram resultados significativos ($p<0,05$), em comparação com 97% nos estudos originais, demonstrando um declínio substancial dos efeitos observados e o insucesso de replicação na maioria dos casos.

8.5. Obrigações e Boas Práticas

Sumário de boas práticas, por cada fase do processo de investigação

1 - Planeamento / Pré-Registo:

- Definir perguntas de investigação claras, testáveis e delimitadas.
- Registar antecipadamente hipóteses, métodos e plano analítico.
- Garantir amostras de dimensão adequada (cálculo de poder estatístico).
- Definir critérios de inclusão/exclusão antes da recolha de dados.
- Escolher instrumentos validados e protocolos padronizados.
- Antecipar riscos éticos e preparar medidas de proteção de dados.
- Criar estrutura de pastas e convenções de nomes antes de começar.

- O planeamento sólido é a base da reprodutibilidade.

2 - Recolha de Dados:

- Documentar em detalhe todos os procedimentos
- Registar alterações efetuadas durante a recolha (logbook).
- Utilizar ferramentas de recolha consistentes e auditáveis
- Armazenar dados brutos de forma segura, imutável e com backups.
- Registar metadados completos (quem, quando, como, com que instrumento).
- Registar versões dos questionários / instrumentos utilizados.
- Minimizar erros através de validações automáticas, duplas entradas ou QC.
- Sem documentação rigorosa da recolha a replicação é impossível.

3 - Processamento e Limpeza de Dados:

- Separar dados brutos de dados limpos (pastas / versões distintas).
- Documentar cada transformação (data cleaning log).
- Automatizar processamento com scripts (R, Python) -evitar ações “manuais”.
- Utilizar controlo de versões (Git/GitHub/GitLab).
- Criar dicionários de dados e códigos (codebooks) completos.
- Verificar outliers, entradas inconsistentes e dados em falta.
- Garantir rastreabilidade: qualquer passo deve poder ser refeito por terceiros.
- A transparência nesta fase é um grande fator de reprodutibilidade.

4 - Análise de Dados:

- Utilizar scripts totalmente reproduzíveis.
- Documentar todas as decisões analíticas (p.ex. modelos, parâmetros).

- Registar versão de software e bibliotecas utilizadas.
- Evitar p.ex. p-hacking, HARKing e a exploração não declarada de hipóteses.
- Utilizar pipelines automatizados e ambientes controlados.
- Validar modelos com testes independentes e métricas de robustez.
- Guardar outputs intermédios e logs para auditoria.
- A reprodutibilidade depende da capacidade de repetir todos os passos de análise.

5 - Documentação e Gestão de Projeto:

- Criar README completo para orientar outros utilizadores.
- Manter registo de auditoria de todas as alterações (versões).
- Adotar convenções de nomes claras e padronizadas.
- Documentar dependências, requisitos e ambiente de computação.
- Produzir diagramas de fluxo ou workflow dos processos.
- Guardar todas as decisões metodológicas e justificações.
- Atualizar documentação à medida que o projeto evolui.
- "Se não está documentado, não existe" – princípio da ciência reprodutível.

6 - Publicação e Transparência:

- Seguir diretrizes existentes nas áreas disciplinares relevantes.
- Explicitar métodos totalmente, não apenas resumir.
- Descrever limitações e análises exploratórias.
- Incluir fluxogramas e anexos que ajudem replicadores.
- Evitar omitir dados que "não confirmam" hipóteses.
- Declarar conflitos de interesse e financiamento.

- Usar a CRediT Taxonomy do NISO, ou similar.

7 - Partilha de Dados e Código:

- Depositar dados e código em repositórios FAIR.
- Atribuir DOI/PID aos conjuntos de dados e scripts.
- Preparar versões anónimas ou sintéticas para dados sensíveis.
- Partilhar código executável, documentação e ambientes (Binder, Docker).
- Usar licenças abertas claras (CC-BY, MIT, GPL).
- Garantir preservação a longo prazo (formatos abertos e estáveis).
- Sem partilha, a ciência não pode ser verificada nem reutilizada.

8 - Replicação e Auditoria:

- Replicar internamente antes de publicar (independent analyst replication).
- Incentivar replicações externas e estudos de robustez.
- Publicar ficheiros de replicação completos.
- Criar repositórios com instruções de execução passo a passo.
- Participar em redes de replicação.
- Tratar replicações negativas como oportunidades de melhoria dos processos de investigação, e da documentação para replicação.
- A replicação é parte do processo científico, não é um ataque ao investigador/a.

Tema	Práticas Obrigatórias	Boas Práticas Opcionais (Recomendadas)
Gestão de Dados	Gestão de dados em conformidade com os princípios FAIR.	Utilizar os princípios CARE (Collective Benefit, Authority to Control, Responsibility, Ethics) para dados de comunidades.
Documentação / Validação	Fornecer informação detalhada sobre quaisquer resultado de investigação ou ferramentas ou instrumentos necessários para reutilizar ou validar os dados.	Usar Jupyter Notebooks e eNotebooks para documentação de proveniência e de processos usados. Usar ficheiros README para descrever dados e código.
Software e Métodos	O PGD deve descrever a estratégia para disponibilizar as ferramentas e/ou softwares necessárias para o acesso, uso e reutilização dos dados.	Partilhar o código-fonte em repositórios abertos (e.g., Zenodo/GitHub). Utilizar licenças Open Source (e.g., MIT, GPL) para software.
Reprodutibilidade	Documentação completa, a disponibilização de informação sobre as ferramentas e instrumentos necessários para validar os dados	Usar ferramentas como Binder (mybinder.org) para criar ambientes computacionais reprodutíveis (Validação). Registo de *workflows* no WorkflowHub.
Integridade	Cumprimento das diretrizes de Integridade Científica (COPE) e a transparência na autoria, que está ligada à CRediT Taxonomy	Fazer o pré-registo (Pre-registration) de estudos ou publicar Registered Reports para garantir a transparência da metodologia.
Metadados	Metadados devem ser FAIR e legíveis por máquina, incluindo PID e Proveniência.	Utilizar templates de PGD que suportam PGDs legíveis por máquina (maDMPs) (e.g., ARGOS).

9. Princípios de Gestão de Dados de Investigação e DMPs

9.1. Conceitos Chave

O PGD (DMP) é um documento obrigatório para financiadores como a FCT e Horizonte Europa e deve ser dinâmico e vivo, atualizado ao longo do ciclo de vida do projeto. É um documento formal que descreve a proveniência, organização e curadoria, acesso, preservação, partilha e eventual eliminação dos outputs da investigação. Uma versão curta é necessária na fase de proposta. O PGD da FCT está alinhado com o modelo da Science Europe e recomenda o uso do ARGOS (OpenAIRE) para maDMPs. O DMP serve para o investigador refletir sobre riscos, como lacunas em backups ou proteção de dados pessoais. A versão final, que descreve como os dados são geridos e partilhados, tem de ser entregue no final do projeto. Os custos associados à RDM/GDI – Gestão de Dados de Investigação (curadoria, staff time, encargos de repositório) são elegíveis para financiamento no Horizonte Europa. Os custos devem ser estimados e incluídos no orçamento do projeto, pois são elegíveis apenas durante a duração do mesmo. O PGD deve ser registado como um deliverable público não restrito (exceto quando justificado).

O modelo da Science Europe define 6 Core Requirements (CRs) que cobrem os seguintes aspetos: descrição dos dados, documentação, metadados, armazenamento, segurança, legal, ética, partilha, preservação, responsabilidades e recursos.

9.2. Princípios Fundamentais do RDM e Requisitos para PGDs

Gestão de Dados de Investigação (RDM)

A Gestão de Dados de Investigação (GDI) abrange todo o ciclo de vida de um estudo de investigação, desde o planeamento, geração/recolha de dados, análise, documentação, preservação até à reutilização. A GDI é uma prática de apoio fundamental para assegurar a qualidade e a reutilização dos dados, sendo orientada pelos princípios FAIR, e uma GDI correta ajuda a garantir que os dados seguem estes princípios. Todas estas práticas e requisitos devem ser formalmente apresentados no PGD.

Requisitos Essenciais para um PGD

Os requisitos fundamentais que um PGD deve cobrir são seis:

1. Informação sobre os dados de investigação: Descrição dos dados e condições de reutilização.
2. Documentação e metadados: Especificação dos padrões de metadados a serem utilizados.
3. Armazenamento e segurança: Medidas durante o projeto.
4. Aspetos legais e éticos: Conformidade com o RGPD, Propriedade Intelectual (PI), MoUs/Non-Disclosure Agreements (NDAs).
5. Partilha e preservação a longo prazo: Escolha do repositório, licença e Identificadores Persistentes (PIDs).
6. Responsabilidades e recursos: Definição de custos, PI, infraestrutura e o papel do Data Steward (Gestor de Dados).

Detalhes sobre a Informação e Qualidade dos Dados (Secções 1 e 2)

O PGD deve detalhar o tipo, formato e volume dos dados gerados. É crucial dar preferência a formatos abertos e padrão (como CSV ou Parquet), justificando a necessidade de usar formatos fechados apenas se for imprescindível. Se houver reutilização de dados existentes, a origem e as condições de acesso e reutilização (licenças) devem ser descritas. O PGD deve indicar os standards de metadados e vocabulários controlados aplicáveis (gerais e específicos do domínio científico). Deve ser detalhada a organização da documentação, por exemplo, a estrutura de ficheiros e pastas. Além disso, devem ser descritas as medidas de controlo de qualidade, como calibração, medições repetidas ou validação de entrada de dados. O PGD é um documento vivo e não um documento estático. Não há respostas absolutamente certas, desde que as decisões sejam claras, específicas e detalhadas, com justificação.

Armazenamento, Segurança e Legal (Secções 3 e 4)

O PGD deve descrever onde e como os dados serão armazenados durante o projeto, bem como a política de backup (tipo e frequência). Para dados sensíveis, devem ser descritas medidas de segurança adicionais, como encriptação e acesso restrito, e as políticas

institucionais de proteção de dados.

O PGD deve assegurar explicitamente o cumprimento do RGPD para dados pessoais, incluindo a garantia do consentimento informado para a preservação e partilha futura. A anonimização dos dados pessoais deve ser considerada, pois dados verdadeiramente anónimos já não são considerados dados pessoais. Para projetos de colaboração, as questões de PI e os direitos de controlo de acesso devem ser definidos e cobertos pelo Acordo de Consórcio. No caso de partilha restrita (por NDA ou para proteger dados pessoais), o PGD deve justificar a razão para a restrição. As justificações para não abrir dados devem separar claramente razões legais e contratuais de restrições intencionais.

Partilha, Preservação e Recursos (Secções 5 e 6)

O PGD deve detalhar que dados serão preservados e onde, sendo preferível um repositório certificado. A FCT exige a preservação por pelo menos 10 anos. Se as condições do concurso exigirem, o repositório deve ser federado no EOSC. A licença de reutilização deve ser indicada, sendo que os metadados devem ser sempre abertos e sob licença CC0 (Domínio Público), e devem ser acessíveis mesmo que os dados não estejam disponíveis. O repositório deve atribuir Identificadores Persistentes (PIDs), como o DOI, para cada conjunto de dados.

A estratégia para disponibilizar as ferramentas e software necessários para aceder e reutilizar os dados (incluindo a sustentabilidade do software) deve ser descrita. Os outputs de investigação, ferramentas e instrumentos necessários para a validação das conclusões da publicação podem incluir software, algoritmos, protocolos, modelos, workflows e electronic notebooks. O PGD deve delinear quem é o responsável pela GDI (PI, Data Steward) e que recursos (financeiros e tempo) serão dedicados à curadoria dos dados e à garantia do FAIR. É importante notar que os custos de armazenamento e curadoria são elegíveis para programas de financiamento como o Horizonte Europa.

RDM e Ciência Aberta no Horizonte Europa – Obrigações Principais

O Horizonte Europa impõe obrigações rigorosas para a Ciência Aberta: Publicações Científicas:

- Acesso aberto a publicações científicas deve ser garantido, depositando a cópia eletrónica da versão publicada ou o manuscrito final revisto por pares (AAM) num repositório trusted, no máximo, à data da publicação.

- Deve ser usada a versão mais recente da licença CC BY ou equivalente. Para

monografias ou textos longos, é possível excluir o uso comercial e obras derivadas. Capítulos em livros editados são tratados de forma similar a artigos de revista para efeitos de licenciamento, não se qualificando para as licenças restritivas de textos longos.

- Os beneficiários (ou autores) devem manter direitos de PI suficientes para cumprir os requisitos de acesso aberto.

- Os metadados das publicações depositadas devem estar abertos sob CC0. Estes metadados devem incluir informações sobre a publicação, o financiamento (Horizon Europe ou Euratom), o projeto, os termos de licenciamento e PIDs para a publicação, autores, organizações e resultados/ferramentas necessárias para a validação.

- Apenas são elegíveis as taxas de publicação em espaços de acesso aberto total para publicações científicas com revisão por pares.

Dados de Investigação:

- Os beneficiários devem gerir os dados digitais de investigação em conformidade com os princípios FAIR.

- É obrigatório definir e atualizar regularmente um Plano de Gestão de Dados (PGD).

- Os dados devem ser depositados num repositório confiável o mais rapidamente possível e dentro dos prazos estabelecidos no PGD. Se as condições do concurso o exigirem, o repositório deve ser federado na EOSC.

- O acesso aberto deve ser garantido através do repositório, utilizando a versão mais recente do CC BY ou CC0 ou uma licença com direitos equivalentes.

- O princípio a seguir é "o mais aberto possível, tão fechado quanto necessário". O acesso aberto pode ser negado se for contrário aos interesses legítimos do beneficiário (incluindo exploração comercial), contrário a restrições (segurança, dual use) ou a obrigações contratuais. Se o acesso não for aberto, a justificação deve constar no PGD, e as restrições devem ser incluídas nos metadados.

- Em caso de emergência pública, o acesso aos dados deve ser dado imediatamente, sob CC BY, CC0 ou equivalente.

- Adicionalmente, os beneficiários devem fornecer acesso (digital ou físico) aos dados, ferramentas e instrumentos necessários para a validação das conclusões das publicações científicas, salvaguardando-se os interesses legítimos ou restrições.

Requisitos Detalhados da Science Europe para PGDs (6 Core Requirements)

O guia da Science Europe detalha os requisitos essenciais que devem ser abordados no PGD:

1. Descrição dos dados e recolha ou reutilização de dados existentes:

- Como serão recolhidos ou produzidos novos dados e/ou como serão reutilizados os dados existentes (metodologias e software). Deve ser explicada como a proveniência dos dados será documentada.

- Que dados (tipo, formatos e volumes) serão recolhidos ou produzidos. Os formatos devem ser justificados, dando-se preferência a formatos abertos e padrão (ex: .txt, .rdf) para facilitar a partilha e reutilização a longo prazo. Os volumes podem ser expressos em espaço de armazenamento (bytes) ou em número de objetos/ficheiros.

2. Documentação e qualidade dos dados:

- Que metadados e documentação (metodologia de recolha, organização dos dados) acompanharão os dados. Devem ser indicados os metadados para ajudar na descoberta, quais os padrões de metadados a utilizar (ex: DDI, TEI, EML, MARC, CMDI) e como os dados serão organizados (controlo de versões, estruturas de pastas). A documentação deve incluir definições de variáveis e unidades de medida, e o local onde a informação será capturada (readme file, codebooks, cadernos de laboratório).

- Que medidas de controlo de qualidade serão utilizadas, como calibração, medições repetidas, captura padronizada, validação de introdução de dados, revisão por pares de dados ou representação com vocabulário controlado.

3. Armazenamento e backup durante o processo de investigação:

- Como serão armazenados e copiados os dados e metadados durante a investigação, e com que frequência. Recomenda-se o armazenamento em pelo menos dois locais separados, dando-se preferência ao uso de armazenamento

robusto e gerido com backup automático fornecido pela instituição (desaconselhando-se pen drives ou discos rígidos autónomos).

- Como será assegurada a segurança dos dados e a proteção de dados sensíveis, incluindo como os dados serão recuperados em caso de incidente. Deve ser explicado quem terá acesso aos dados e como é controlado, especialmente em parcerias colaborativas.

4. Requisitos legais e éticos, códigos de conduta:

- Se dados pessoais forem processados, como será garantido o cumprimento da legislação (RGPD) e segurança. Isto inclui obter consentimento informado para preservação/partilha, considerar a anonimização ou pseudonimização (como a encriptação) e explicar se existe um procedimento de acesso gerido para utilizadores autorizados de dados pessoais.

- Como serão geridas outras questões legais, como os direitos de propriedade intelectual (PI) e a propriedade. Deve-se explicar quem será o proprietário dos dados (quem terá os direitos de controlo de acesso), as condições de acesso (aberto ou restrito) e como serão tratados os direitos sui generis da Base de Dados. Estas questões devem ser cobertas no acordo de consórcio para projetos colaborativos.

- Que questões éticas e códigos de conduta serão tidos em conta. Deve-se demonstrar consciência de como as questões éticas afetam o armazenamento, transferência e acesso aos dados. É necessário verificar se a revisão ética (por um comité de ética) é exigida para a recolha de dados.

5. Partilha de dados e preservação a longo prazo:

- Como e quando os dados serão partilhados, e se existem restrições ou razões de embargo. Deve-se explicar como os dados serão descobertos (repositório confiável, indexados em catálogo, tratamento direto de pedidos). Deve ser definido o plano de preservação (quanto tempo serão mantidos). Se o uso exclusivo for reivindicado ou a partilha adiada (ex: para PI ou patentes), deve ser justificado.

- Como serão selecionados os dados para preservação e onde serão preservados a longo prazo (repositório ou arquivo de dados). Deve ser indicado quais os dados a

reter ou destruir e onde serão depositados. Recomenda-se demonstrar que o repositório é estabelecido e que as suas políticas (metadados, custos) foram verificadas.

- Que métodos ou ferramentas de software são necessários para aceder e utilizar os dados. Deve ser considerada a sustentabilidade do software.

- Como será garantida a aplicação de um Identificador Persistente (PID), como um DOI, a cada conjunto de dados. Os PIDs permitem que os dados sejam localizados e referenciados de forma fiável, e ajudam a acompanhar citações.

6. Responsabilidades e recursos de gestão de dados:

- Quem (função, cargo e instituição) será responsável pela gestão de dados (data manager). Devem ser descritos os papéis e responsabilidades (p.ex., captura de dados, produção de metadados, backup, arquivo). Para projetos colaborativos, deve ser explicada a coordenação das responsabilidades entre os parceiros.

- Que recursos (financeiros e de tempo) serão dedicados à GDI e para garantir que os dados serão FAIR. Devem ser explicados os recursos necessários para a curadoria de dados (tempo de pessoal, armazenamento, hardware, encargos de repositório) e como esses custos serão cobertos.

Alinhamento do PGD com o Ciclo de Vida do Projeto

O PGD deve ser um documento dinâmico e vivo, alinhado com o Plano de Gestão do Projeto (PMP) em todas as fases:

Fase do Projeto	Foco do PMP (Plano do Projeto)	Foco do PGD (Gestão de Dados)	Pontos de Ligação e Dependência Mútua
1. ANÁLISE E CONCEÇÃO (Ideia, Proposta, Candidatura)	Âmbito (Scope) e Objetivos: Definir os objetivos de investigação, os Work Packages (WPs)e Deliverables (entregáveis).	PGD Inicial (DMP v1.0): Definir os data-outputs chave (quais os dados e o seu volume), a abordagem de Open Science (FAIR e *As Open As Possible, As Closed As Necessary*).	Decisões Estratégicas: O PMP (WPs) define o que a equipa vai fazer; o PGD define quais dados serão gerados pelas tarefas.
2. PLANEAMENTO (Definição da Metodologia)	Recursos, Tempo e Riscos: Atribuição de tarefas (Responsabilidades - PMP), cronograma (Milestones), e planos de mitigação de riscos (ex: falhas de equipamento, perda de pessoal) e orçamento.	Organização e Ética: Definir a metodologia de recolha, os formatos (padrões), as normas de metadados e os procedimentos de ética e consentimento (RGPD).	Interdependência: O PMP deve garantir que as responsabilidades do PGD (Gestor de Dados/Data Steward) estão formalmente atribuídas (Secção 6 do PGD). Os riscos no PMP (ex: perda de dados) dependem dos procedimentos de backup do PGD (Secção 3/5 do PGD). O budget do PMP deve alocar recursos para o PGD (Responsabilidades e Custos).
3. EXECUÇÃO (Implementação e Recolha de Dados)	Controlo e Monitorização: Gestão de equipas, progresso de WPs, controlo de qualidade e gestão de alterações (mudanças de âmbito, metodologia ou prazos).	Armazenamento e Curadoria: Implementação dos procedimentos de segurança e backup (armazenamento ativo). Curadoria de Dados ativa (limpeza, validação, controlo de qualidade e versionamento).	Garantia de Qualidade: O PMP verifica se o PGD está a ser executado. O PGD (Garantia de Qualidade) é um input crucial para o Controlo de Qualidade do PMP. A necessidade de atualizar o DMP (ex: DMP v2.0) é um evento de gestão de alterações do PMP.
4. FECHO DO PROJETO (Conclusão e Arquivo)	Fecho Administrativo: Entrega de todos os Deliverables (incluindo o DMP final), relatório final, e fecho do budget.	Disseminação e Preservação: Seleção e depósito final dos dados no repositório. Atribuição de Identificadores Persistentes (PID). Aplicação da Licença de Acesso Aberto e formalização da preservação a longo prazo.	Fecho Legal/Técnico: O projeto só pode terminar após o PGD Final ter sido aprovado e os dados terem sido depositados de forma FAIR e em conformidade legal. O PMP valida que os custos de preservação foram cobertos.

9.3. Estrutura e Ferramentas para PGDs

Os modelos de PGDs da FCT, Horizon Europe (HE) e European Research Council (ERC) são muito similares nas suas categorias essenciais (Informação sobre Dados, Documentação/Metadados, Armazenamento/Segurança, Dados Pessoais/PI, Partilha/Preservação, Responsabilidades/Recursos).

Categoria Principal do PGD	Horizon Europe (HE	European Research Council (ERC)	FCT
1. Informação sobre Dados	**"1. Data Summary"**	**"SUMMARY"**	**"1. INFORMAÇÃO SOBRE DADOS"**
	- Reutilização de dados existentes (e motivos se descartada).	- Nome do conjunto de dados.	- Que dados existentes serão reutilizados?
	- Tipos, formatos e tamanho esperado.	- Origem e tamanho esperado.	- Que dados serão gerados?
	- Propósito e relevância para os objetivos.	- Tipos e formatos de dados.	- Tipos, formatos e dimensão dos dados.
2. Documentação e Metadados	**"2.1 Making data findable, including provisions for metadata"**	**"1. MAKING DATA FINDABLE"**	**"2. DOCUMENTAÇÃO E METADADOS"**
	- Uso de identificadores persistentes (PID).	- Descrição detalhada do dataset.	- Tipo de documentação (ficheiro README, diários de laboratório).
	- Criação de rich metadata.	- Uso de metadados, PID (DOI, etc.).	- Padrões de metadados a utilizar.
	- Padrões de metadados disciplinares/gerais.		- Identificadores persistentes (DOI) e a sua atribuição.

Categoria Principal do PGD	Horizon Europe (HE	European Research Council (ERC)	FCT
3. Armazenamento e Segurança	**"5. Data security"**	**"5. ALLOCATION OF RESOURCES and DATA SECURITY"**	**"4. ARMAZENAMENTO E SEGURANÇA"**
	- Medidas para segurança de dados (recuperação, armazenamento seguro, transferência de dados sensíveis).	- Procedimentos para backup e recuperação.	- Procedimentos de backup e recuperação durante o projeto.
	- Uso de repositórios fiáveis para preservação a longo prazo.	- Transferência de dados sensíveis e armazenamento seguro.	- Solução de armazenamento e segurança a longo prazo.
4. Dados Pessoais, PI e Propriedade	**"6. Ethics"**	**Não é uma secção separada no corpo do DMP.**	**"3. ÉTICA, LEGAL E PROPRIEDADE INTELECTUAL"**
	- Ética ou questões legais que afetam a partilha.	- O **"DISCLAIMER"** no final clarifica que a ética não faz parte do DMP, mas o PI deve informar a Ethics Team.	- Dados Pessoais e RGPD (Anonimização, Pseudonimização, Consentimento).
	- Consentimento informado para partilha e preservação.		- Questões de IPR e Propriedade (Direitos de Autor).

Categoria Principal do PGD	Horizon Europe (HE	European Research Council (ERC)	FCT
5. Partilha e Preservação a Longo Prazo	**"2.2 Making data openly accessible" "2.3 Making data interoperable" "2.4 Increasing data re-use"**	**"2. MAKING DATA OPENLY ACCESSIBLE" "3. MAKING DATA INTEROPERABLE" "4. INCREASE DATA RE-USE"**	**"5. PARTILHA E PRESERVAÇÃO A LONGO PRAZO"**
	- Acesso aberto (com justificação para exceções).	- Acesso (Onde, Como, Ferramentas/Software).	- Disponibilização de dados (Abertos vs. Fechados; Justificação).
	- Padrões de vocabulário (Interoperabilidade).	- Padrões e vocabulários (Interoperabilidade).	- Repositório, Licença e Ferramentas/Software necessários.
	- Licenciamento (Reutilização).	- Licenciamento, Embargo, Qualidade de Dados (Reutilização).	- Período de preservação (a partir de quando e por quanto tempo).
6. Responsabilidades e Recursos	**"4. Resources"**	**"5. ALLOCATION OF RESOURCES and DATA SECURITY"**	**"6. RESPONSABILIDADES E RECURSOS"**
	- Recursos necessários (Custos, Pessoal).	- Custos estimados para acesso aberto.	- Quem (função, cargo) será responsável pela gestão de dados?
	- Quem é o responsável pela gestão de dados.	- Valor da preservação a longo prazo.	- Recursos (financeiros e de tempo) dedicados à gestão de dados.
	- Como a preservação a longo prazo será assegurada.		

O ARGOS do OpenAIRE é uma ferramenta open source recomendada que facilita o cumprimento das políticas de Acesso Aberto e dados FAIR. O ARGOS permite a criação de PGDs legíveis por máquina (maDMPs), o que possibilita a automação, a colaboração e o versionamento dos PGDs.

ARGOS é um serviço gratuito e open source. Suporta identificadores persistentes como ORCIDs e DOIs. É interoperável e utiliza o Research Data Alliance DMP Common Standard. O DMP criado pode ser versionado (mantendo o histórico e a proveniência). O PGD pode ser publicado e preservado diretamente no Zenodo. Permite ainda criar templates (dataset profiles) adaptados aos padrões específicos de domínio.

9.4. Casos de Estudo: PGDs Reais de Projetos Europeus

A análise de PGDs reais (ex: CrAFt, UNIFY, REPAIR, AfricanBioServices, TRACE, BENEFIT, DyViTo, ATARCA) demonstra como a estrutura do PGD se adapta a diferentes domínios e requisitos. Os projetos em consórcios (a maioria dos exemplos) abordam a coordenação das responsabilidades de GDI entre parceiros.

Salientam-se a diferença nos PGDs entre o Horizonte 2020 (open data pilot), em que a fase piloto foi para alguns projetos, e os PGDs do Horizonte Europa, em que as obrigações de CA/GDI são para todos os projetos, com modelos e regras mais maturados.

Os PGDs diferem conforme o domínio e os dados específicos, como o caso do ATARCA (HE) que lida com dados transacionais de blockchain (públicos/pseudónimos) e dados de tokenization, não recolhendo dados pessoais sensíveis (raça ou saúde). O projeto TRACE (ERC), na área da saúde, exige que o armazenamento dos dados biomédicos seja feito em servidor dentro da firewall do Karolinska Institutet e que o acesso aos dados clínicos subjacentes seja apenas mediante pedido. O PGD do LP-NORM (ERC) foca-se na preservação de dados numéricos de robótica em formatos text-based (CSV, Compressed Sparse Column) para garantir a interoperabilidade e evitar discrepâncias em floating-point numbers. O LP-NORM armazena o software no GitHub e utiliza o Zenodo para a preservação a longo prazo.

Abaixo são apresentados exemplos de vários programas, destacando-se as diferenças entre os modelos seguidos, devido a diferentes orientações e obrigações por parte de cada financiador.

Projetos de Ciências Naturais (p.ex. AfricanBioServices) focam-se em standards de metadados geoespaciais e biológicos (p.ex. dados 2GIS e DRSRD aerial surveys) e na reutilização de dados de levantamentos anteriores, define nove variáveis comuns obrigatórias (geográficas, temporais e de propriedade/autoria) para a recolha de dados de campo, e utiliza uma plataforma de repositório interna como plataforma ativa para armazenamento e troca de todos os dados para o consórcio.

Projetos de Saúde (p.ex. TRACE) exigem acesso gerido (managed access) para garantir a proteção de dados (RGPD).

Projetos de Engenharia (p.ex. ATARCA) focam-se na interoperabilidade de dados e podem exigir DPIA para Apps.

Projetos de Ciências Sociais (p.ex. BENEFIT) focam-se no consentimento informado e no acesso controlado para proteger os participantes.

Os modelos de PGD podem ser adaptados para uma simplificação, como nos casos de programas como ERC (p.ex. TRACE) e MSCA (p.ex. CyViTo).

Exemplos de PGDs reais, selecionados no CORDIS, por área temática:

Estudos Urbanos e Sustentabilidade (Arquitetura, Design Participativo, Cidades Inteligentes - NEB) – Projeto CrAFt - DMP. Astronomia e Astrofísica – Projeto LSP-MIST - DMP.

Ciência de Materiais e Física Experimental (Difração de Raio-X, Medições Ferroelétricas) –

Projeto UNIFY - DMP.

Robótica e Otimização Numérica (Sistemas de Controlo Robótico, Otimização de Precisão, Benchmarks Numéricos)– Projeto LP- NORM - DMP.

Ciência Climática e Ambiental (Emissões de gases não-CO2, Modelação Climática e Simulações) – Projeto REPAIR - DMP.

Ciências da engenharia e tecnologias – Projeto ATARCA - PGD. Ciências naturais – Projeto AfricanBioServices - PGD.

Ciências médicas e da saúde – Projeto TRACE - PGD.

Ciências sociais – Projeto BENEFIT - PGD.

Ciências agrárias – Projeto Circular Agronomics - PGD.

Humanidades – Projeto DyViTo - PGD.

9.5. Obrigações e Boas Práticas

Tema / Área	Práticas Obrigatórias (Mandatory) (HE/ERC/FCT)	Boas Práticas Opcionais (Recomendadas)
DMP / Planeamento	PGD obrigatório como deliverable. Atualização regular (documento vivo).	Utilizar ferramentas que suportam PGDs machine-actionable (e.g., ARGOS). Tornar o PGD público.
Estrutura / Conteúdo	Cobrir as 6 secções principais (dados, documentação, segurança, legal/ética, partilha, recursos). Fornecer um resumo na proposta.	Publicar o PGD no Zenodo com DOI para versionamento. Incluir o costing tool do OpenAIRE para justificar custos de RDM.
Conformidade Legal	PGD deve detalhar o cumprimento do RGPD para dados pessoais. Justificar no PGD o fecho ("as closed as necessary").	Utilizar ferramentas de anonimização (e.g., AMNESIA) para permitir a partilha responsável. Formalizar DSAs (Acordos de Partilha de Dados) em consórcios.
FAIR Data	Dados devem ser FAIR. Usar PIDs (DOI) para datasets e Trusted Repositories.	Usar Vocabulários Controlados e Standards Comunitários (consultar Fairsharing.org). Incluir código/software (com licença Open Source) para validação.
Recursos e Responsabilidade	PGD deve detalhar recursos (tempo/financeiro) dedicados ao FAIR.	Nomear formalmente o Data Steward responsável pelo PGD e pela qualidade dos dados.

9.6. Modelos de PGD para a ULisboa

9.6.1. Modelo de PDG completo

Este modelo de Plano de Gestão de Dados (PGD) é sugerido: trata-se de uma adaptação para a Universidade de Lisboa (ULisboa), fundindo os requisitos obrigatórios do Horizonte Europa (HE), ERC e FCT com as melhores práticas de Ciência Aberta.

Recomenda-se a utilização da ferramenta ARGOS (OpenAIRE) para a criação e atualização deste PGD no formato legível por máquina (maDMP), facilitando a colaboração, o versionamento e a conformidade.

Plano de Gestão de Dados (PGD) – ULisboa

Versão: [V1.0 - Inicial / V2.0 - Intercalar / V3.0 - Final] Data: [DD/MM/AAAA]

Identificador Persistente do PGD (DOI): [Atribuído pelo repositório após a publicação do PGD]

Informação Administrativa

Título do Projeto: [Título Completo do Projeto]
Acrónimo: [Acrónimo do Projeto]

Organização Financiadora: [FCT / Comissão Europeia (HE/ERC) / Outra] N.º da Bolsa/Contrato: [Número do Contrato]

Investigador(a) Principal (PI): [Nome e ORCID do PI] Organização Coordenadora: Universidade de Lisboa Parceiros: [Lista de Instituições Parceiras (se aplicável)]

Contacto (Data Steward/Gestor de Dados): [Nome, Função, Email]

1. Informação sobre os Dados (Tipos, Formatos e Volume)

Esta secção descreve que dados serão gerados e/ou reutilizados no projeto, com ênfase na escolha de formatos que garantam a interoperabilidade e a preservação a longo prazo.

1.1. Dados a Gerar e Proveniência: Serão recolhidos ou produzidos os seguintes tipos de dados (e.g., numéricos, textuais, imagem, áudio, vídeo). Os formatos de dados a serem utilizados serão [especificar formatos e volumes], devendo ser dada preferência a formatos abertos e padrão (ex: CSV, PDF/A, ODT, Parquet), justificando-se o uso de formatos proprietários apenas se for imprescindível. A proveniência dos dados (histórico de alterações, metodologias e software utilizados) será documentada no PGD e nos metadados.

1.2. Reutilização de Dados Existentes: Caso se reutilizem dados de fontes existentes, deve-se descrever a sua origem (referência), formato e as condições de acesso e reutilização (licenças). Caso a reutilização tenha sido considerada e descartada, as razões devem ser brevemente expostas.

2. Documentação e Qualidade dos Dados (Metadados e Reprodutibilidade)

Esta secção garante que os dados serão compreensíveis e reutilizáveis por terceiros (Princípios F e R de FAIR).

2.1. Padrões de Metadados: Os metadados devem ser ricos para permitir a descoberta dos dados. Será utilizado o Dublin Core como padrão de metadados mínimo, e serão adotados os seguintes padrões específicos do domínio [indicar padrões, p.ex., DDI, EML, consultando Fairsharing.org]. Os metadados devem estar abertos sob licença CC0 (Domínio Público) para maximizar a descoberta e a colheita (harvesting).

2.2. Documentação de Suporte: A documentação para garantir a reutilização incluirá a metodologia de recolha, as definições de variáveis, unidades de medida e a organização dos dados. Será criado um ficheiro README e/ou codebook (dicionário de dados) para contextualizar o dataset. A documentação será capturada e mantida em [especificar local, p.ex., cadernos de laboratório digitais, Jupyter Notebooks, cabeçalhos de ficheiros].

2.3. Controlo de Qualidade: As medidas para garantir a consistência e qualidade da recolha de dados serão [descrever medidas, e.g., calibração de instrumentos, medições repetidas, validação de dados de entrada, revisão por pares de dados ou vocabulário controlado].

3. Armazenamento e Segurança dos Dados (Durante a Investigação)

Esta secção descreve como os dados serão protegidos contra perdas e acesso não autorizado durante a fase ativa do projeto.

3.1. Armazenamento Ativo e Backup: Durante a investigação, os dados e metadados serão armazenados em [especificar local, p.ex., servidores de rede da ULisboa ou cloud institucional]. Deve ser dado preferência a armazenamento robusto e gerido com backup automático. A política de backup será [detalhar tipo e frequência], e os dados serão armazenados em pelo menos dois locais separados para evitar perdas. O armazenamento em dispositivos externos (como pen drives ou discos rígidos autónomos) não é recomendado.

3.2. Segurança e Proteção de Dados Sensíveis: A segurança será assegurada através de [descrever medidas de segurança, p.ex., controlo de acesso via credenciais institucionais]. Para dados sensíveis (p.ex., pessoais), serão aplicadas medidas de segurança adicionais, como a encriptação. Será explicado como os dados serão recuperados em caso de incidente.

4. Aspetos Legais, Éticos e Propriedade Intelectual (PI)

Esta secção aborda o cumprimento do RGPD, a ética e o enquadramento legal do dataset.

4.1. Conformidade com o RGPD e Dados Pessoais: Se dados pessoais forem processados, será garantido o cumprimento do RGPD (Lei n.º 58/2019). É obrigatório obter consentimento informado para a preservação e/ou partilha futura dos dados pessoais. Para partilha, será considerada a anonimização robusta para que os dados deixem de ser considerados pessoais ou, em alternativa, a pseudonimização (e.g., encriptação). Deve ser explicado se existe um procedimento de acesso gerido para utilizadores autorizados de dados pessoais e de como o acesso aos dados é controlado em parcerias.

4.2. Uso de Inteligência Artificial (IA): O investigador é sempre responsável pelo conteúdo gerado por IA, e a autoria não pode ser atribuída à IA. É proibida a introdução de dados críticos (dados pessoais, confidenciais ou protegidos por PI/Direitos de Autor) em ferramentas GenAI abertas ou não aprovadas pela instituição, devido ao risco de fuga de informação e incumprimento do RGPD. O uso substancial de ferramentas de IA deve ser divulgado de forma transparente, incluindo o nome e a versão do modelo, e o método de utilização. O PGD deve garantir a preservação dos registos (logs) de atividade dos sistemas de IA para fins de auditoria e rastreabilidade. No caso de sistemas de alto risco, é crucial documentar a origem dos dados de treino e as medidas para mitigar enviesamentos (bias).

4.3. Propriedade Intelectual (PI): [Nome da Instituição/Parceiros] será o proprietário dos dados (detentor dos direitos de controlo de acesso). Para projetos colaborativos, estas

questões estão definidas no Acordo de Consórcio. Qualquer restrição de acesso devido a PI ou segredos comerciais deve ser justificada.

4.4. Ética e Códigos de Conduta: Serão seguidos os códigos de conduta e diretrizes éticas nacionais e internacionais. Será verificada a necessidade de revisão ética para a recolha de dados.

5. Partilha e Preservação a Longo Prazo Esta secção define o plano de disseminação (FAIR) e a preservação sustentável dos dados (Princípios A e R).

5.1. Princípio de Abertura e Restrições: Os dados serão disponibilizados sob o princípio "o mais aberto possível, tão fechado quanto necessário". Se os dados forem restritos (e.g., por RGPD, PI ou segurança da UE), esta justificação deve ser incluída nos metadados. O acesso a dados restritos será feito através de um procedimento de acesso gerido formalizado por Acordos de Partilha de Dados (DSAs).

5.2. Repositório e PIDs: Os dados serão depositados num repositório confiável (e.g., Repositório Institucional da ULisboa, Zenodo, ou repositório disciplinar certificado CoreTrustSeal) o mais rapidamente possível. Será garantida a atribuição de um Identificador Persistente (PID), como o DOI, a cada conjunto de dados. A FCT exige a preservação por, pelo menos, 10 anos.

5.3. Licenciamento: A licença de reutilização para os dados depositados será a versão mais recente de CC BY ou CC0.

5.4. Ferramentas para Validação: Serão disponibilizados os dados e outros resultados, ferramentas e instrumentos necessários para validar as conclusões das publicações e reutilizar os dados. Isto inclui software, algoritmos e protocolos. A estratégia para disponibilizar o software (com licença Open Source) será [detalhar, p.ex., via Zenodo/GitHub]. Recomenda-se o uso de ferramentas como o Binder para criar ambientes computacionais reprodutíveis.

6. Responsabilidades e Recursos de Gestão de Dados

Esta secção aloca responsabilidades e recursos para garantir o cumprimento do PGD e os princípios FAIR.

6.1. Papéis e Responsabilidades: O Investigador Principal (PI) é o responsável final pelo cumprimento do PGD. [Nome, Função, Instituição] será o Data Steward/Gestor de Dados

responsável pela execução das atividades de GDI (p.ex., curadoria, produção de metadados, backup).

6.2. Recursos e Custos: Os recursos de tempo e financeiros dedicados à Gestão de Dados (p.ex., tempo de pessoal para curadoria, custos de armazenamento e encargos de repositório) serão [descrever custos e como serão cobertos]. É importante notar que os custos de armazenamento e curadoria são elegíveis para programas de financiamento como o Horizonte Europa.

6.3. Rastreabilidade e Auditoria: Os registos (logs) da atividade de IA ou outras decisões críticas de gestão de dados serão preservados para fins de auditoria e rastreabilidade (accountability), o que é uma responsabilidade de gestão.

9.7. Modelo de PDG simplificado

Este é um modelo de Plano de Gestão de Dados (PGD) simplificado e mínimo, desenhado para projetos internos da ULisboa ou experiências de pequena escala, com foco no cumprimento dos requisitos obrigatórios de conformidade e nos princípios FAIR.

Plano de Gestão de Dados (PGD) Simplificado – ULisboa

Versão: V1.0
(Inicial) Data:
[DD/MM/AAAA]

Identificador Persistente do PGD (DOI): [Registo a ser feito no repositório institucional no final do projeto, se aplicável]

Informação Administrativa

Título do Projeto / Estudo: [Título do Estudo ou Projeto Interno]
Investigador(a) Principal (IP): [Nome do IP/Responsável]

Unidade Orgânica / Centro de Investigação: [Nome da Unidade ULisboa]
Organização Financiadora: [ou: Projeto Interno / Sem Financiamento Externo]

Contacto Principal (para questões de dados): [Nome e Email do IP ou Gestor de Dados]

1. Informação sobre os Dados (Tipos, Formatos e Volume)

Os dados gerados ou recolhidos no projeto serão de [especificar tipos de dados, e.g., numéricos, textuais, de imagem] e o volume esperado é de [estimar volume, e.g., 5 GB, 200 ficheiros]. Deve- se dar preferência a formatos abertos e padrão (e.g., CSV, PDF/A, TXT) para garantir a interoperabilidade e a preservação a longo prazo. Se forem reutilizados dados existentes, a sua origem, formato e licença de uso devem ser documentados.

2. Documentação e Qualidade dos Dados (Metadados)

2.1. Metadados: Serão aplicados os metadados mínimos (e.g., Dublin Core) e os metadados específicos do domínio, se aplicáveis. Os metadados descritivos dos dados devem estar abertos sob licença CC0 (Domínio Público).

2.2. Documentação: A documentação para permitir a reutilização dos dados incluirá [especificar, e.g., um ficheiro README, um livro de códigos (codebook) ou notas de laboratório] que descrevem a metodologia de recolha, as definições de variáveis e as unidades de medida. Serão aplicadas medidas de controlo de qualidade (e.g., calibração, validação de entrada de dados).

3. Armazenamento e Segurança dos Dados (Durante a Investigação)

3.1. Armazenamento: Os dados ativos e metadados serão armazenados no(s) sistema(s) de rede seguro(s) da ULisboa, ou em soluções institucionais de cloud gerida, sendo desaconselhado o uso de dispositivos de armazenamento externos como pen drives ou discos rígidos autónomos.

3.2. Segurança e Backup: Deve ser utilizada uma solução de backup robusta e gerida, com os dados armazenados em pelo menos dois locais separados. Serão implementadas medidas de segurança como controlo de acesso e, para dados confidenciais ou sensíveis, encriptação, para garantir a proteção e recuperação em caso de incidente.

4. Aspetos Legais, Éticos e Propriedade Intelectual (PI)

4.1. Conformidade com o RGPD: Caso sejam processados dados pessoais ou sensíveis, é obrigatório o cumprimento rigoroso do Regulamento Geral de Proteção de Dados (RGPD). Deve ser obtido o consentimento informado dos participantes para a recolha e, se aplicável, para a preservação e partilha futura dos dados. Para partilha, os dados

pessoais serão submetidos a anonimização robusta (dados verdadeiramente anónimos) ou a pseudonimização (e.g., encriptação com chave separada). Em casos sensíveis, a revisão ética por um comité é obrigatória.

4.2. Uso de IA: O investigador é o único responsável pela integridade do output, e a autoria não pode ser atribuída à IA. É proibida a introdução de dados críticos (pessoais, confidenciais ou protegidos por PI/Direitos de Autor) em ferramentas GenAI abertas ou não aprovadas pela instituição. Se a IA for utilizada, deve ser feita uma declaração obrigatória sobre o seu uso para transparência.

4.3. Propriedade Intelectual: A [Unidade Orgânica / ULisboa] será a proprietária dos dados (detentora dos direitos de controlo de acesso).

5. Partilha e Preservação a Longo Prazo

5.1. Princípio de Abertura: Os dados serão disponibilizados segundo o princípio "tão aberto quanto possível, tão fechado quanto necessário". Se houver restrições de acesso (e.g., por RGPD, ou segredo comercial/PI), tal deve ser justificado e as restrições devem ser incluídas nos metadados.

5.2. Repositório e Preservação: Os dados finais a preservar serão depositados num repositório confiável (e.g., Repositório Institucional da ULisboa ou Zenodo). Será atribuído um Identificador Persistente (PID), como o DOI, a cada conjunto de dados. O período de preservação dos dados será de [definir período, e.g., 10 anos (padrão FCT) ou mais, dependendo do domínio].

5.3. Licenciamento: A licença de reutilização aplicada ao dataset depositado será CC BY ou CC0.

5.4. Ferramentas para Validação: Serão disponibilizados os métodos, software ou ferramentas necessárias para aceder, validar ou reutilizar os dados.

6. Responsabilidades e Recursos

6.1. Responsabilidades: O Investigador Principal (IP) é o responsável pela implementação do PGD e por garantir que este é revisto e atualizado.

6.2. Recursos: Os recursos (tempo de pessoal e hardware) dedicados à gestão e curadoria dos dados serão assegurados no âmbito da [Unidade Orgânica].

10. Utilização de Inteligência Artificial (IA) em gestão de dados e investigação

10.1. Conceitos Chave

A Inteligência Artificial (IA) está a transformar todo o ciclo de investigação (planeamento, análise, automação de código), desde a conceção de ideias até à publicação, utilizando modelos de linguagem (LLMs) e IA generativa (GenAI) para agilizar o planeamento de projetos, a análise de dados, a automatização de código e o apoio à redação científica, o que pode aumentar a qualidade da produção científica. A IA pode ser usada na investigação, mas exige a máxima transparência e rastreabilidade. O investigador é sempre o responsável final por qualquer output gerado pela IA, incluindo a verificação de alucinações/citações falsas. A autoria não pode ser atribuída à IA.

A IA é definida como um sistema baseado em máquina que opera com diferentes níveis de autonomia e que, a partir do input recebido, infere como gerar outputs como predições ou conteúdo. A IA generativa (GenAI) aprende padrões a partir de dados de treino e prevê o próximo passo, seja na criação de texto, imagem ou som. Os Large Language Models (LLMs) são essencialmente máquinas de correspondência de padrões (pattern matching machines) e não sistemas capazes de raciocínio lógico. Esta natureza resulta em que os LLMs tendem a gerar uma média de todo o conteúdo a que foram expostos e, consequentemente, podem replicar todos os enviesamentos (bias) e erros existentes nos dados de treino. O uso de LLMs pode levar à desinformação e a atos de má conduta como a fabricação, falsificação e plágio.

O seu uso é regido pelo AI Act da UE, que estabelece quatro níveis regulamentares de risco (inaceitável, alto, limitado (como os LLMs), e mínimo). Os projetos de IA usados exclusivamente em I&D ("*in the lab*") têm uma isenção legal.

A IA é uma ferramenta poderosa na GDI, auxiliando p.ex. na geração automática de metadados, na anonimização inteligente e na otimização de workflows. Especificamente na GDI, a IA pode automatizar a extração de metadados ricos de ficheiros, acelerando a curadoria de dados. Além disso, algoritmos de Machine Learning e LLMs podem acelerar a pseudonimização e anonimização de grandes volumes de texto ou voz, contribuindo para a conformidade ética e do RGPD. O PGD deve ser complementado com uma descrição da IA usada, mitigação de viés (bias) e rastreabilidade (logs) para garantir a integridade e transparência. O uso substancial de ferramentas de IA generativa deve ser

detalhado de forma transparente (p.ex. nome da ferramenta/modelo, versão, como foi usada). O Acesso Aberto é um combustível fundamental para as ferramentas de IA de nova geração. O *Text and Data Mining* (TDM) só é possível em larga escala quando os artigos estão em formatos abertos e legíveis por máquina (como XML ou JSON) sob licenças que permitam uso automático dos dados. Sem o Acesso Aberto, os modelos de linguagem seriam treinados apenas em conteúdos proprietários ou de baixa qualidade, limitando as capacidades das ferramentas de IA.

É recomendado proibir a introdução de dados críticos e sensíveis (p.ex. pessoais, confidenciais, PI) em ferramentas de AI não aprovadas pela instituição. Outras recomendações de fiabilidade incluem ter atenção a, e mitigar, enviesamentos causados pelos dados de treino.

10.2. IA no Ciclo de Investigação e Implicações Éticas/Legais

Enquadramento da IA

É importante notar que AI é diferente de ML, que é diferente de um LLM. A IA é uma tecnologia relativamente antiga, com a primeira referência da ideia a surgir em 1950 (Alan Turing) e o termo "Inteligência Artificial" a ser formalmente usado em 1955 (John McCarthy). Os primeiros modelos, por exemplo, para reconhecimento de padrões, foram criados por volta de 1956/58. A IA utiliza uma abordagem estatística e os seus resultados não são determinísticos, pois dependem do desenho da rede neuronal, dos dados de treino e do contexto de utilização.

Transformação das Atividades de Investigação pela IA

A IA, sobretudo a IA generativa, pode ajudar em todas as fases do ciclo de vida da investigação. Isto inclui a fase de planeamento (p.ex., melhorando queries de pesquisa), a criação de documentação (p.ex., readme files), até à análise de dados, onde pode gerar código-fonte de interpretação (p.ex., em linguagens como Python). No entanto, todos estes drafts gerados pela IA devem ser submetidos a revisão e edição humana.

As grandes vantagens da IA incluem melhorias no acesso, como a tradução, a melhoria de textos em outras línguas e a realização de tarefas de programação, mesmo para investigadores sem conhecimentos de programação. A IA permite acelerar o

brainstorming, a revisão de literatura, a sugestão de gaps de investigação e o desenho experimental, e ainda colaborar na redação de manuscritos. LLMs são particularmente eficazes na automatização de tarefas rotineiras, administrativas e objetivas, como resumir, compilar, categorizar e agrupar informações públicas.

As ferramentas de IA vieram para ficar e estão a mudar a forma de trabalho, mas a adoção é uma barreira substancial e existem riscos significativos. As tarefas que envolvem pesquisar, compilar, resumir, categorizar ou agrupar informação pública, que são rotineiras, administrativas e objetivas, podem ser facilmente automatizadas por LLMs, o que pode poupar tempo.

IA no Ciclo de Investigação (Fases)

A IA auxilia em todas as fases do ciclo de investigação (Planear, Recolher, Processar, Analisar, Preservar, Partilhar e Reutilizar), sendo o Comportamento Responsável na Investigação (RCR) o seu núcleo, abrangendo a Ética e a Lei:

Fase de Investigação	Transformação pela IA	Exemplos de Ferramentas/Métodos de IA	Implicações Principais
1. ANÁLISE DA IDEIA E CONCEÇÃO	Geração de Hipóteses* Criativas e Síntese de Literatura: IA pode processar muitos artigos, identificar lacunas, tendências e relações não óbvias entre diferentes campos para sugerir novas direções de investigação.	LLMs para revisão sintética (e.g., resumir artigos em minutos). Criar gráficos de Conhecimento (Knowledge Graphs) ou **mindmaps** para mapeamento de conceitos.	Velocidade de análise acelerada. Reduz o tempo de revisão de literatura. Risco de enviesamento Algorítmico: O PGD deve agora considerar a origem e o enviesamento dos dados de treino usados pela IA para gerar hipóteses e resumos.
2. PLANEAMENTO E METODOLOGIA	Otimização do Desenho Experimental: Algoritmos ML preditivos poderm melhorar o planeamento (otimizam a alocação de recursos). Uso de AI na metodologia de análise / investigação do próprio conteúdo de investigação.	ML para otimização de planeamento. Para aprendizagem ativa de temas necessários no Projeto. Usar GenAI para preencher PGDs (texto ou no formato maDMP). Gerar partes do PMP e templates do projeto (atas, minutas).	O PGD deve incluir a versão do modelo de IA usado para o planeamento e previstos na metodologia (transparência e reprodutibilidade).

Fase de Investigação	Transformação pela IA	Exemplos de Ferramentas/Métodos de IA	Implicações Principais
3. EXECUÇÃO, RECOLHA E ANÁLISE DE DADOS	Automação e Análise em Tempo Real: Automação da recolha de dados (sensores/visão computacional), pré-processamento (limpeza e normalização) e análise, mais deteção imediata de padrões ou anomalias.	Visão Artificial (para análise de imagens microscópicas ou satélite). Deep Learning para classificação e segmentação automática. Uso de modelos abertos (*open source AI*) para adaptação ao tópico de investigação. Workflows automatizados.	Aumento do Volume de Dados (Big Data): Necessidade de cloud computing e repositórios escaláveis. Problema da 'Caixa Negra' (Black Box): interpretar o Modelo (XAI). O PGD e o PMP devem alocar recursos para a documentação da AI.
4. FECHO E ARQUIVO	Preservação de Outputs Digitais e Documentação: A IA pode automatizar a extração de metadados ricos dos resultados finais e gerar relatórios sumários dos Deliverables.	LLMs para geração de resumos e documentação final de metadados.	Requisito de Reprodutibilidade Computacional: Não basta preservar os dados; é obrigatório preservar o modelo de IA treinado e o ambiente de execução. Isto pode ser problemático para modelos fechados

Utilização da IA para a Gestão de Dados de Investigação (GDI)

A IA pode otimizar as tarefas de Gestão de Dados de Investigação:

- Geração Automática de Metadados: A IA pode ler ficheiros e inferir automaticamente informações contextuais (data de recolha, instrumento, variáveis), gerando metadados ricos e estruturados alinhados com padrões disciplinares (ex: Dublin Core). Também pode realizar verificações de qualidade sobre metadados já preparados. Isto acelera a Curadoria e melhora a completude/consistência dos dados.

- Identificação de Entidades e Relações: LLMs podem identificar termos-chave e relações dentro da documentação de dados, facilitando a pesquisa online de datasets e projetos relacionados. Isto melhora a Descoberta e a localização de datasets através de pesquisa semântica.

- Anonimização Inteligente: Algoritmos de ML e LLMs podem automatizar e acelerar a pseudonimização e anonimização de grandes volumes de dados de texto ou voz, o que contribui para a Conformidade Ética e RGPD.

- Tradução Automática de Vocabulários: A IA pode mapear termos entre vocabulários controlados (p.ex., termos de saúde de diferentes países), aumentando a Interoperabilidade de datasets globais e ultrapassando barreiras semânticas.

- Auxílio à Criação de maDMPs: Modelos de IA podem ler os requisitos de financiadores (FCT, HE) e gerar o PGD no formato legível por máquina (maDMP), preenchendo secções com base na descrição do projeto ou logs de recolha.

- Interface de Consulta de Dados (Chatbots): LLMs podem criar chatbots para interrogar grandes datasets ou repositórios usando linguagem natural, facilitando o Acesso a utilizadores não técnicos.

- Plataformas de descoberta alimentadas por IA podem ser usadas para otimizar a revisão de literatura e a gestão de referências:

    - Semantic Scholar: Uma plataforma gratuita que utiliza Inteligência Artificial para gerar "TLDRs" (resumos extremamente curtos) em áreas como medicina e biologia, ajudando a filtrar a sobrecarga de informação. Destaca-se também pelo sistema de "Citações Altamente Influentes", que identifica trabalhos com impacto real no desenvolvimento de uma linha de investigação.

    - Scilit: Um agregador de conteúdos em tempo real (mantido pela MDPI) que indexa novas publicações em poucas horas através de análise de big data bibliográfica, oferecendo serviços de alerta de citações e de novos artigos (SciFeed) com base em consultas personalizadas.

    - Dimensions AI: Uma plataforma avançada que, para além de artigos e preprints, inclui na sua base de dados ensaios clínicos, documentos de políticas públicas e patentes. Permite a visualização de redes de colaboração e análises bibliométricas complexas para apoiar decisões baseadas em evidência.

    - Scopus e Web of Science: Embora sejam bases de dados por subscrição, são ferramentas de curadoria profissional que permitem identificar facilmente artigos de Acesso Aberto (identificados com ícones específicos, como o cadeado laranja no Scopus) e rastrear o impacto através de indexação rigorosa de citações.

- Controlo de Qualidade e Limpeza de Dados: Algoritmos de ML podem detetar outliers, inconsistências e erros em datasets, sugerindo correções. Isto melhora a Reutilização, garantindo que apenas dados de alta qualidade são preservados.

- Recriação de Ambientes de Execução: A IA pode auxiliar na criação de containers (ex: Docker) ao analisar o código e os requisitos de software, garantindo a Reprodutibilidade Computacional. Também pode criar dados sintéticos para acompanhar os containers no caso de dados sensíveis que não possam ser partilhados.

- Preservação de Modelos de IA: No caso de modelos abertos, a documentação e preservação dos próprios modelos de ML, incluindo o código de treino e os pesos finais, são outputs valiosos que aumentam a Longevidade da Investigação.

Utilização da IA para GDI	Descrição e Vantagens	Impacto na GDI
Geração Automática de Metadados	Ler ficheiros (imagens, tabelas, logs de instrumentos) e inferir automaticamente informações contextuais (data de recolha, instrumento, variáveis) para gerar metadados ricos e estruturados, alinhados com padrões disciplinares (ex: Dublin Core, padrões específicos do domínio). Ou realizar verificações de qualidade sobre metados já preparados versus conjuntos de dados.	Aceleração da Curadoria: Reduz a carga de trabalho manual por parte dos investigadores. Pode melhorar a completude e consistência dos dados.
Identificação de Entidades e Relações	LLMs podem identificar termos chave, nomes de investigadores e relações dentro da documentação de dados. Pode ainda extrair palavras-chave e facilitar a pesquisa online de dados e de projetos relacionados.	Melhora a Descoberta: Facilita a localização de datasets através de pesquisa semântica, ligando dados a publicações e a mais projetos.
Anonimização Inteligente	Algoritmos e LLMs podem realizar pseudonimizaçãoe anonimização, aplicando modelos de machine learning para identificar e mascarar informações pessoais em grandes volumes de dados de texto ou voz.	Conformidade Ética e RGPD: Automatizae acelera um processo complexo e sensível.

Utilização da IA	Descrição e Vantagens	Impacto na GDI
Tradução Automática de Vocabulários	A IA pode mapear termos de um vocabulário controlado para outro (p.ex: mapear termos de saúde de um país), facilitando a combinação de datasets globais.	Aumenta a Interoperabilidade: Permite que dados de diferentes origens e padrões sejam combinados e analisados, para ultrapassar barreiras semânticas.
Auxílio à Criação de maDMPs	Modelos de IA podem ler os requisitos de financiadores (ERC, HE, FCT) e gerar o PGD no formato legível por máquina (maDMP - machine-actionable DMP), preenchendo secções com base na descrição do projeto e relatórios ou até *logs* de recolha dos dados, ou inquéritos.	Aceleração da preparação dos documentos, incluindo o PGD, e facilitar ainda mais a criação de maDMPs.
Interface de Consulta de Dados (Chatbots)	Utilização de LLMs para criar chatbots que permitem aos utilizadores interrogar grandes datasets ou repositórios usando linguagem natural, sem necessidade de saber p.ex. SQL ou linguagens de consulta complexas.	Facilitar o acesso: Torna os dadose outros resultados mais acessíveis a utilizadores não técnicos (cidadãos, decisores políticos, ou outros investigadores).

Utilização da IA	Descrição e Vantagens	Impacto na GDI
Controlo de Qualidade e Limpeza de Dados	Algoritmos de ML podem detetar automaticamente outliers, inconsistências e erros em grandes datasets, sugerir correções ou sinalizar dados de baixa qualidade.	Melhora a Reutilização: Garante que apenas dados de alta qualidade são preservados, o que aumenta a confiança e o valor da reutilização para terceiros.
Recriação de Ambientes de Execução	A IA pode auxiliar na criação de containers (ex: Docker) ao analisar o código e os requisitos de software (versões de bibliotecas, sistemas operativos), para garantir que os modelos de IA e a análise de dados são reprodutíveis. Criação de dados sintéticos para acompanhar os containers, nos casos de processamento de dados sensíveis que não possam ser partilhados.	Reprodutibilidade Computacional: Essencial para a Ciência Aberta. Permite que qualquer pessoa execute a análise original, replicando ou reproduzindo os resultados.
Preservação de Modelos de IA	No caso de modelos abertos, a possibilidade de documentar e preservar os próprios modelos de Machine Learning (o código de treino, os pesos finais e os dados de validação), que são eles próprios outputs valiosos da investigação.	Longevidade da Investigação: Os modelos de IA (o "conhecimento" aprendido) tornam-se objetos de dados preserváveis e citáveis. Compatibilidade de modelos abertos com os princícios de CA.

Riscos e Implicações Éticas

Responsabilidade e Autoria: Um ponto fundamental é que a autoria não pode ser atribuída à IA. Os investigadores são sempre responsáveis por qualquer output gerado (mesmo que por IA), incluindo potenciais más práticas como fabricação, falsificação e plágio. O investigador é o responsável final pelo conteúdo gerado, devendo verificar a exatidão, validade e conformidade com a PI e corrigir quaisquer erros ou inconsistências, especialmente em candidaturas a financiamento e em entregáveis de projetos financiados.

Transparência e Rastreabilidade: É crucial promover a transparência e a rastreabilidade (explainability) do uso de IA, mitigando a opacidade dos modelos e o viés algorítmico. A rastreabilidade através de registos (logs) é uma responsabilidade de gestão fundamental para a integridade. O uso substancial de ferramentas de IA generativa deve ser detalhado (nome, versão, data e método de utilização). Os prompts e os outputs devem estar disponíveis em conformidade com os princípios da Ciência Aberta, se aplicável.

Riscos de Informação Crítica: Existe um risco crítico de fuga de informação. Nunca se deve inserir dados pessoais, informações confidenciais, credenciais de acesso ou ideias-chave de investigação em qualquer LLM aberto. A introdução de dados nestas ferramentas de terceiros pode levar a que os dados sejam usados no treino futuro do modelo ou armazenados nos seus servidores, sem consentimento nem transparência, o que pode configurar um incumprimento do RGPD. Este risco é particularmente relevante com plataformas baseadas fora do Espaço Económico Europeu (EEE), como as sediadas nos EUA, devido à menor proteção do RGPD.

Falha do Pensamento Crítico e Imprecisão: O uso de LLMs pode incentivar a uma "delegação cognitiva" na ferramenta de IA, levando à dependência excessiva, e à falha na revisão crítica dos detalhes, p.ex. devido a falta de tempo ou atenção. Os riscos incluem a obtenção de texto incorreto, citações enganosas (alucinações), e a falta de rastreabilidade da fonte, o que compromete a transparência. São fundamentais os processos de validação de referências e a combinação com processos habituais de *desk research*.

Alguma investigação recente demonstrou que os LLMs têm limitações em problemas complexos e em cálculos exatos, podendo dar resultados de forma inconsistente mesmo em tarefas de média complexidade, e/ou falhas contextuais. Em tarefas de alta complexidade, os modelos podem experienciar um colapso completo.

10.3. O Enquadramento Legal Europeu (AI Act)

O objetivo central do AI Act é garantir que os sistemas de IA no mercado europeu sejam seguros, transparentes, não discriminatórios e respeitem os direitos fundamentais.

Nível de Risco	Descrição	Implicação na Investigação
A. Risco Inaceitável (Proibido)	Sistemas que manipulam o comportamento humano para causar dano ou sistemas de *social scoring* por autoridades públicas. Isto inclui IA que utiliza vulnerabilidades de um grupo específico para causar danos.	Proibição total de desenvolvimento, teste ou utilização, mesmo em ambiente de investigação.
B. Alto Risco (Obrigações Rigorosas)	Potencial significativo para prejudicar a saúde, segurança ou direitos fundamentais (ex: diagnóstico médico, triagem de emprego).	Devem cumprir obrigações rigorosas antes de serem colocados no mercado.
C. Risco Limitado (Obrigações de Transparência)	Sistemas que requerem transparência para que o utilizador tome decisões informadas (ex: Chatbots/LLMs e Deepfakes/conteúdo sintético).	Os investigadores devem informar os utilizadores que estão a interagir com uma IA ou que o conteúdo foi artificialmente gerado.
D. Risco Mínimo ou Nulo	Muitos sistemas de IA (jogos, filtros de spam, geração de metadados internos).	Sem obrigações legais, mas recomenda-se um código de conduta voluntário e práticas de transparência

O AI Act geralmente não se aplica a sistemas desenvolvidos exclusivamente para fins de investigação e desenvolvimento (R&D), antes de serem colocados no mercado. Esta isenção só se aplica enquanto a IA se mantiver como pura investigação (in the lab), mas a aplicabilidade do AI Act entra em vigor quando o projeto se transforma num produto e passa para a operação (aplicação em cenários reais que afetem pessoas). Portanto, ao passar para a operação (aplicação em cenários reais que afetam pessoas), a aplicabilidade do AI Act entra em vigor, podendo classificar o sistema como de alto risco.

As atividades de investigação que usem modelos de propósito geral (GPAI) de alto impacto (como GPT-4 ou Gemini 2.5) têm a obrigação de transparência, gestão de riscos e data governance (transparência sobre os dados de treino).

Um novo desafio ético identificado é o "Alucinamento de Citações". Os investigadores devem validar manualmente todas as referências bibliográficas geradas por LLMs, uma vez que estas ferramentas tendem a criar DOIs e títulos de artigos que parecem plausíveis, mas que não existem. Além disso, para garantir a transparência, qualquer manuscrito que tenha utilizado IA generativa deve incluir uma declaração de utilização de IA, detalhando qual a ferramenta usada (ex: GPT-4o) e para que tarefa específica (p.ex: tradução, revisão de texto, ou otimização de código).

10.4. Privacidade, Conformidade (RGPD) e Integridade com a IA

Conformidade com o RGPD

O RGPD deve ser rigorosamente cumprido em todas as aplicações de IA que processem dados pessoais. O foco na privacidade implica a proibição absoluta de introdução de dados críticos (pessoais, confidenciais ou protegidos por copyright) em ferramentas GenAI não aprovadas pela instituição.

Para dados sensíveis, o uso em IA (para treino ou entrada de dados como p.ex. em perguntas, "*prompts*") exige a anonimização ou pseudonimização dos dados de entrada. Recomenda-se o uso de ferramentas como AMNESIA para anonimização robusta, antes de enviar os dados para a IA. O PGD deve agora descrever não só a origem dos dados de treino, mas também as medidas para mitigar e documentar enviesamentos (bias) nos datasets. A Anonimização Robusta, de preferência com ferramentas como AMNESIA, é

crucial antes de usar os dados como input para a IA, garantindo a conformidade com o RGPD. O foco na mitigação de enviesamentos (bias) deve incluir a análise da paridade e a sub-representação de grupos (p.ex., minorias) nos dados de treino, pois isso afeta o desempenho do modelo.

Transparência e XAI

A documentação deve ser abrangente, e o PGD deve ter a preservação dos registos de atividade (logs) do sistema de IA (decisões, alterações) para auditoria e rastreabilidade (accountability). Esta rastreabilidade (logs) é essencial para a transparência e para a gestão de riscos, especialmente em sistemas de Alto Risco. Para sistemas de IA de alto risco, os metadados no PGD devem ser extremamente detalhados para permitir que as autoridades e os utilizadores compreendam o seu funcionamento (princípio da Interpretabilidade – XAI, eXplainable AI).

Mitigação de Enviesamentos e Qualidade

É necessário detetar enviesamentos nos dados de treino e implementar validações e métricas de equidade. Por exemplo, deve-se analisar a sub-representação de grupos (p.ex., minorias) no dataset de treino, pois isso leva a um pior desempenho do modelo para esses grupos. Métodos como retrieval-augmented generation (RAG) podem melhorar a fiabilidade dos outputs ao minimizar enviesamentos nos dados de treino. O Controlo de Qualidade e Limpeza de Dados pela IA garante que apenas dados de alta qualidade são preservados, o que melhora a Reutilização.

Políticas Institucionais e Diretrizes

As instituições devem definir ferramentas de IA aprovadas e não aprovadas, medidas de salvaguarda, e o tipo de informação crítica que não pode ser inserida na IA. A metodologia de IA deve ser prevista nas políticas e, em casos sensíveis, autorizada por um comité de ética.

As orientações da Comissão Europeia sobre o Uso Responsável da IA Generativa na Investigação (ERA Forum) complementam o AI Act e enfatizam os seguintes princípios:

1. Fiabilidade: Garantir a qualidade, precisão e reprodução da informação gerada pela IA, com atenção a viés e imprecisões. O investigador deve manter-se responsável e abordar criticamente os outputs da IA, verificando a precisão e validade de citações.

2. Honestidade: Divulgar o uso de IA generativa de forma transparente e imparcial.

3. Respeito: Considerar o impacto social, a privacidade, a confidencialidade e os direitos de Propriedade Intelectual. Deve-se ter cautela ao fornecer dados sensíveis ou não publicados à IA, pois estes podem ser usados para treino, violando o RGPD ou a PI.

4. Responsabilidade: O investigador é o único responsável pela produção científica, e os sistemas de IA não são autores nem coautores.

Os sistemas de IA desenvolvidos em investigação devem ser tecnicamente robustos (precisos e reprodutíveis) e socialmente robustos (considerando o contexto social e minimizando danos), sendo capazes de fornecer uma explicação adequada do seu processo de tomada de decisão quando têm um impacto significativo em pessoas.

10.5. Casos de Estudo

Exemplos de Aplicações Práticas

Apoio à Redação e Pesquisa: Ferramentas como NotebookLM e ChatPDF permitem conversar com documentos carregados, fornecendo texto sintetizado e referências citáveis para validação. Outros LLMs (Gemini, Claude, ChatGTP) podem ser usados para as etapas iniciais de desk research. A utilização de APIs académicas, como a Semantic Scholar Academic Graph API, permite que investigadores extraiam dados estruturados sobre citações e autores em escala, possibilitando a integração de algoritmos de aprendizagem automática (machine learning) personalizados para analisar tendências de investigação sem as limitações das interfaces de utilizador tradicionais. O NotebookLM tem a capacidade de gerar um podcast sobre o artigo carregado e criar mindmaps para facilitar a compreensão, além de fornecer referências citáveis.

Comunicação Científica: A IA pode adaptar abstracts para comunicação num estilo mais acessível para uso em redes sociais ou gerar infografias simples e imagens ilustrativas de conceitos investigados a partir de descrições simples (ex: DALL-E, Gemini).

Aplicações Científicas Robustas: O AlphaFold é um exemplo de avanço científico em IA, que decifrou 200 milhões de estruturas de proteínas, embora se tenha baseado em 200.000 estruturas previamente decifradas através de investigação humana de 50 anos.

Projetos de Alto Risco: O projeto AI-Mind (H2020) ilustra a necessidade de managed access e de conformidade rigorosa com o RGPD, pois utiliza algoritmos de deep learning para analisar dados clínicos sensíveis (EEG, MRI) de 1000 participantes. O PGD deste tipo de projeto deve detalhar a ética e a trustworthiness das técnicas de IA utilizadas.

Má Conduta em Casos de Estudo: Investigadores enfrentaram o problema de paper mills (fábricas de papers) e publicações falsas geradas por IA, que já foram alvo de trabalhos de investigação. Também foram descobertos casos em que os LLMs geraram citações falsas associadas a autores existentes.

10.6. Obrigações e Boas Práticas

Para a preparação de candidaturas a financiamento (como no Horizonte Europa), os proponentes são totalmente responsáveis pelo conteúdo gerado pela IA e devem ser transparentes ao divulgar quais ferramentas foram usadas e como. É responsabilidade dos proponentes verificar a precisão e validade do conteúdo e das citações geradas pela IA e corrigir quaisquer erros ou inconsistências. Além disso, deve-se verificar as fontes originais para garantir a conformidade com as regulamentações de Propriedade Intelectual e evitar plágio.

Os sistemas baseados em IA devem ser desenvolvidos para serem tecnicamente robustos, precisos e reprodutíveis, para lidar com falhas, e serem socialmente robustos, considerando o contexto em que operam. Devem também ser capazes de fornecer uma explicação adequada do seu processo de tomada de decisão, se tiverem um impacto significativo na vida das pessoas.

Tema / Área	Práticas Obrigatórias	Boas Práticas Opcionais (Recomendadas)
Integridade e Autoria	IA não pode ser reconhecida autoria de um resultado. O investigador é responsável pela realização do trabalho e por más praticas eventuais como fabricação, falsificação e plágio. Declaração obrigatória sobre o uso de IA em investigação e publicações.	Utilizar a IA de forma explicável (*explainability*), mantendo registos claros (*logs*) dos modelos utilizados para fins de auditoria.
Privacidade (RGPD)	Cumprimento rigoroso do RGPD em todas as aplicações de IA que processem dados pessoais. Aplicação de anonimização ou pseudonimização para dados sensíveis.	Proibir o input de dados críticos (pessoais, material confidencial/sensível ou dados protegidos por copyright) em ferramentas GenAI não aprovadas pela instituição. Utilizar AMNESIA para anonimização robusta antes de input para AI.
Enquadramento Legal / Ético	A metodologia e os processos de IA devem ser autorizados para investigação em matérias sensíveis. Distingir o uso de IA em investigação pura vs. operação (alto risco/AI Act).	Colaboração multidisciplinar (legal, ética, técnica) para o desenvolvimento de soluções de IA.
PGD e Rastreabilidade	O PGD deve incluir definições claras sobre o uso de IA, especificando modelos usados, validação e auditoria.	Incluir a estratégia de licenciamento e proveniência para dados de treino de IA no PGD.

Anexo I – Indicadores de sucesso sugeridos

A definição de Indicadores de Sucesso para monitorizar a adesão e o impacto da política de Ciência Aberta (CA) na gestão e partilha de dados de investigação na ULisboa deve focar-se em métricas que indiquem a transição dos investigadores para práticas mais abertas, transparentes e reprodutíveis, conforme os requisitos de financiadores como o Horizonte Europa (HE) e a FCT, e as diretrizes internacionais (FAIR e CoARA).

Esta seção apresenta uma lista de indicadores que podem ser usados pela ULisboa para medir adesão e impacto. Não deve ser considerada como uma lista para implementação completa, mas como um conjunto de possibilidades, para escolha da melhor forma de monitorização, consoante o contexto. Os indicadores de sucesso podem ser categorizados por tipo de output e pela fase do ciclo de investigação:

I. Adesão à Política de Acesso Aberto (Publicações)

Estes indicadores medem o cumprimento dos mandatos de acesso imediato e o uso de licenças abertas para publicações:

1. Taxa de Acesso Aberto Imediato (AA) em Repositórios Confiáveis:

- Métrica: Percentagem de artigos científicos revistos por pares (Author Accepted Manuscript - AAM ou Version of Record - VoR) depositados num repositório confiável (como o Repositório ULisboa ou Zenodo).
- Requisito de Adesão: O depósito deve ocorrer, o mais tardar, na data de publicação (immediate OA).

2. Conformidade do Licenciamento de Publicações:

- Métrica: Percentagem de publicações depositadas que utilizam a licença CC BY 4.0 ou equivalente, assegurando que os autores retiveram direitos suficientes (Rights Retention Strategy).
- Indicador de Qualidade: Número de monografias ou textos longos que utilizam licenças CC mais restritivas (CC BY-NC/ND) quando devidamente justificado.

3. Adoção de Plataformas de Publicação Transparente:

- Métrica: Número de publicações de investigadores da ULisboa submetidas à

plataforma Open Research Europe (ORE), que é gratuita, Diamond Open Access, e implementa Open Peer Review.

II. Qualidade e Abertura dos Dados de Investigação (FAIRness)

Estes indicadores medem a diligência dos investigadores na gestão e documentação dos dados, em conformidade com o princípio "tão aberto quanto possível, tão fechado quanto necessário".

1. Partilha de Dados Abertos e Licenciamento:

- Métrica: Número e percentagem de datasets finais e agregados publicados em repositórios confiáveis (trusted repositories).
- Requisito de Adesão: Os datasets abertos devem ser licenciados sob CC BY ou CC0.

2. Qualidade dos Metadados (Findable e Accessible):

- Métrica: Percentagem de datasets publicados com metadados abertos (CC0), detalhados (rich metadata) e machine-actionable.
- Indicador de Adesão: Uso consistente de Identificadores Persistentes (PIDs) para os dados (DOI) e os autores (ORCID).

3. Utilização do Plano de Gestão de Dados (PGD):

- Métrica: Número de projetos que produzem dados que entregam o PGD (Data Management Plan) como deliverable, seguindo os requisitos obrigatórios (e.g., modelo FCT/HE).
- Indicador de Qualidade: Uso de ferramentas de PGD machine-actionable (maDMPs), como o ARGOS, para facilitar a interoperabilidade e a automação.

4. Transparência Metodológica (Partilha Abrangente):

- Métrica: Contagem de outputs de investigação adicionais (software, algoritmos, protocolos, modelos, workflows, electronic notebooks) disponibilizados em repositórios abertos (e.g., GitHub, Zenodo, WorkflowHub) para validar as conclusões.
    - Métrica de Apoio: Número de itens de research software publicados com

SWHID e/ou com licença open source.

- Indicador de Adesão: Presença e qualidade da Data Availability Statement (DAS) em publicações, citando explicitamente os PIDs e as condições de acesso aos dados.

III. Indicadores de Processo, Ética e Impacto (Adoção Qualitativa)

Estes indicadores focam-se nas boas práticas e na integração da CA na cultura de investigação, em linha com a reforma da avaliação (CoARA):

1. Integridade e Reprodutibilidade:

- Métrica: Número de estudos submetidos a pré-registo (pre-registration) ou publicados como relatórios registados (registered reports).

- Indicador de Qualidade: Uso de ferramentas de ambientes computacionais reprodutíveis (e.g., Binder, Jupyter Notebooks) para documentar a proveniência e a análise dos dados.

2. Conformidade Legal e Ética:

- Métrica: Número de projetos que envolvem dados pessoais que documentam a conformidade com o RGPD (e.g., DPIA ou Revisão Ética), e que utilizam ferramentas de anonimização (e.g., AMNESIA) antes da partilha.

- Indicador: Partilha de dados sensíveis sob acesso controlado (managed access) formalizado por Acordos de Partilha de Dados (DSAs), garantindo o Authority to Control (Princípios CARE).

3. Engajamento e Colaboração:

- Métrica: Número de projetos que utilizam a Citizen Science (Ciência Cidadã).

- Indicador de Qualidade: Adoção da taxonomia CRedIT Taxonomy para reconhecer formalmente as contribuições dos colaboradores (incluindo curadoria de dados e software) em publicações.

4. Reutilização e Impacto:

- Métrica: Monitorização do uso e impacto dos outputs abertos através de

altmetrics (e.g., downloads, menções em redes sociais, reutilização dos datasets).

- Indicador Estratégico: Percentagem de relatórios de avaliação de carreira/institucional que reconhecem e valorizam a qualidade e diversidade dos outputs abertos (dados, software, protocolos) em detrimento de métricas quantitativas de publicação (Impact Factor).

Anexo II – Sumário de obrigações e boas práticas

Obrigações

- Seguir princípios FAIR e garantir preservação por 10 anos.
- Usar identificadores persistentes (DOI, ORCID obrigatório) e formatos abertos.
- Publicar só em trusted repositories. Um repositório confiável deve ter mecanismos para garantir a precisão, integridade e autenticidade dos seus conteúdos.
- Sem períodos de embargo para publicações em acesso aberto.
- É obrigatório fornecer Acesso Aberto imediato (sem embargo) a artigos científicos (AAM ou VoR) sob licença CC BY 4.0 ou equivalente ou o machine-readable electronic copy da versão publicada.
- Os Metadados (de dados de acesso aberto) devem ser FAIR, acessíveis e licenciados sob CC0 (Domínio Público).
- O PGD (Plano de Gestão de Dados) é obrigatório como deliverable, deve ser dinâmico (atualização regular em caso de mudanças significativas) e cobrir as 6 secções principais (dados, documentação, segurança, legal/ética, partilha, recursos).
- O PGD deve detalhar a alocação de recursos e custos (tempo/financeiro) dedicados ao FAIR e à preservação dos dados.
- O PGD deve definir a propriedade dos dados (direitos de controlo de acesso) e a gestão da PI.
- O PGD deve detalhar o cumprimento do RGPD para dados pessoais. Justificar no PGD o fecho ("as closed as necessary").
- Cumprir a legislação sobre dados pessoais (RGPD) e garantir a proteção e segurança dos dados sensíveis, e integrar as questões de proteção de dados no PGD.
- Obter consentimento informado (livre, específico, informado e inequívoco) para a

recolha, preservação e/ou partilha de dados pessoais.

- Fornecer uma Notificação de Privacidade (Privacy Notice) clara antes de começar a recolha de dados.
- Garantir Backup e armazenamento seguro, encriptar dados confidenciais.
- Realizar uma Avaliação de Impacto sobre a Proteção de Dados (DPIA/AIPD) para o tratamento de dados de alto risco, e manter registos de pedidos e respostas (princípio da accountability).
- Fornecer informação detalhada sobre qualquer resultado de investigação ou ferramentas (softwares, algoritmos, protocolos, modelos, workflows e electronic notebooks) ou instrumentos necessários para reutilizar ou validar as conclusões.
- O Investigador deve aplicar o Dublin Core mínimo obrigatório e standards de domínio, adotar formatos abertos (CSV, PDF/A), usar convenções de nomeação com data ISO e redigir um README.md com a documentação.
- Submeter a metodologia a Revisão Ética (Comité de Ética) em todos os casos aplicáveis.
- Cumprir as diretrizes de Integridade Científica (COPE) e a transparência na autoria, que está ligada à CRediT Taxonomy.
- Metadados devem ser FAIR e legíveis por máquina, incluindo PID e Proveniência.
- O processo de anonimização ou pseudonimização aplicado aos dados deve ser explicitamente detalhado no PGD.
- IA não pode ser reconhecida autoria de um resultado. O investigador é responsável pela integridade dos outputs (e por más praticas eventuais como fabricação, falsificação e plágio). Exigir declaração obrigatória do uso de IA em investigação e publicações.
- A metodologia e os processos de IA devem ser autorizados para investigação em matérias sensíveis.
- O PGD deve incluir definições claras sobre o uso de IA, especificando modelos usados, validação e auditoria.

Boas Práticas

- Recomenda-se utilizar ferramentas PGD que suportam a automação (maDMPs), como o ARGOS, e que o PGD seja público.
- Publicar o PGD no Zenodo com DOI para versionamento, e incluir o costing tool do OpenAIRE para justificar custos de RDM.
- Depositar em repositórios que possuam Certificação CoreTrustSeal para garantir a preservação a longo prazo.
- Usar mais do que um repositório, e buscar a preservação por mais de 10 anos.
- A Estratégia de Retenção de Direitos deve ser utilizada preventivamente em todas as submissões a revistas.
- Recomenda-se publicar em revistas full Open Access (Gold OA), cujos custos são elegíveis, mas as APCs para revistas híbridas não são elegíveis.
- Utilizar licenças CC mais restritivas (CC BY-NC/ND) só devem ser usadas quando justificadas (ex: monografias ou para proteger interesses comerciais).
- Licenciar dados brutos (raw data) com CC0 ou CC BY resolve dúvidas sobre direitos conexos e maximiza a abertura.
- Usar Vocabulários Controlados e Standards Comunitários (consultar Fairsharing.org).
- Adoção de standards de domínio (ex: Darwin Core, EML, SWHID) através dos Data Stewards.
- Utilizar PIDs para todos os elementos (software, protocolos) e ROR para instituições.
- Associar dados, publicações, e software: datasets, artigos e código ligados com PIDs (DOI, ORCID, ROR) e descrições cruzadas nos metadados.
- Partilhar outros resultados (código, algoritmos, protocolos, workflows) necessários para a validação.
- Partilhar o código-fonte em repositórios abertos (e.g., Zenodo/GitHub). Utilizar licenças Open Source (e.g., MIT, GPL) para software.

- Usar ferramentas de criação de metadados legíveis por máquinas (e.g., CEDAR).
- Usar Jupyter Notebooks e eNotebooks para documentação de proveniência e de processos usados. Usar ficheiros README para descrever dados e código.
- Usar ferramentas como Binder (mybinder.org) para criar ambientes computacionais reprodutíveis (Validação). Registo de workflows no WorkflowHub.
- Fazer o pré-registo (Pre-registration) de estudos ou publicar Registered Reports para garantir a transparência da metodologia.
- A Data Availability Statement (DAS) deve ser adicionada no final do artigo antes da submissão.
- Monitorizar reutilização e impacto (p.ex., downloads, citações e reutilização dos dados –

  altmetrics).
- Realizar Avaliações de Impacto sobre a Proteção de Dados (AIPDs) para refletir sobre os riscos e as salvaguardas.
- Envolver o Encarregado de Proteção de Dados (DPO) da instituição em todas as fases do projeto.
- Garantir anonimização ou pseudonimização sempre que possível para reduzir riscos.
- Utilizar AMNESIA (ou similar) para anonimização robusta, retirando os dados do âmbito do RGPD.
- Adotar os Princípios CARE (Collective Benefit, Authority to Control) para dados de comunidades vulneráveis.
- Formalizar Acordos de Partilha de Dados (DSAs) para dados confidenciais ou sensíveis e/ou pseudonimizados.
- Proibir o input de dados críticos (pessoais, confidenciais, PI/Direitos de Autor) em ferramentas GenAI não aprovadas pela instituição. Manter registos claros (logs) do uso de IA para transparência e rastreabilidade.

- Colaboração multidisciplinar (legal, ética, técnica) para o desenvolvimento de soluções de IA.
- Incluir a estratégia de licenciamento e proveniência para dados de treino de IA no PGD.
- Nomear formalmente o Data Steward responsável pelo PGD e pela qualidade dos dados.
- O depósito em plataformas comerciais como ResearchGate, Academia.edu, websites pessoais ou serviços cloud (Dropbox, Google Drive) é explicitamente não conforme e nenhum destes é considerado um Repositório Confiável.