Guia para a Gestão de Dados de Investigação na NOVA FCSH

Enquadramento e objetivo

Os dados de investigação assumem atualmente uma importância comparável às publicações enquanto resultados de I&D, contribuindo para a reprodutibilidade, impacto, transparência e excelência da investigação levada a cabo nas instituições de ensino superior. Neste sentido, as agências financiadoras do Sistema Científico e Tecnológico nacional (STCN), como a FCT e a Comissão Europeia, têm vindo a incluir como requisito para atribuição de financiamento a implementação de práticas de gestão de dados de investigação (GDI) nos projetos, incluindo a elaboração de planos de gestão de dados (PGD) e o depósito dos conjuntos de dados em repositórios confiáveis e em acesso aberto, sempre que possível.

Este documento reúne um conjunto de boas práticas de GDI de forma a auxiliar os/as investigadores/as, docentes, gestores/as de ciência e demais intervenientes na gestão de dados resultantes de projetos de I&D. Estas boas práticas estão organizadas pelas várias fases do ciclo de vida dos dados de investigação, incluindo o planeamento, a recolha e organização, o processamento e análise, a publicação e partilha, a preservação e a reutilização.

A implementação das boas prática de GDI será apoiada pelo Núcleo de Desenvolvimento Digital da Investigação (NDDI) da NOVA FCSH, sobretudo na preparação de PGD e no arquivo de conjuntos de dados de investigação em repositórios confiáveis.

Boas práticas de GDI

Planeamento

- **Planear a investigação.** Na fase de planeamento é importante refletir sobre a produção, preservação e partilha dos dados de investigação no âmbito das Ciências Sociais, Artes e Humanidades (CSAH). Deve identificar-se o tipo de financiamento, entidade financiadora e equipa de investigação, definir-se os requisitos técnicos e fixar-se as etapas da investigação, assegurando o cumprimento dos princípios FAIR (https://www.go-fair.org/fair-principles/) e a eficiência, integridade, transparência e confiabilidade do processo científico (https://doi.org/10.5281/zenodo.3820473).

- **Criar um plano de gestão dos dados (PDG)**. Um PGD é um documento formal e dinâmico, que abrange todas as fases do processo de investigação e pode ser atualizado sempre que necessário. Constitui um dos requisitos-base das agências de financiamento para a atribuição de fundos públicos para projetos científicos. A sua apresentação é obrigatória para projetos com financiamento europeu (programas-quadro Horizonte 2020 e Horizonte Europa) e é aconselhada para projetos financiados pela FCT (Política sobre a disponibilização de dados e outros resultados de projetos de I&D, 2014). O PGD pode ser criado pelo/a investigador/a ou ser desenvolvido a partir de modelos disponibilizados pelas entidades financiadoras, através de ferramentas como o DMPonline e o Argos, devendo contemplar:
    - Uma descrição do projeto (financiamento, sumário, objetivos, equipa);
    - Uma descrição dos conjuntos de dados (*datasets*) gerados e/ou reutilizados: formato, tipologia e dimensão; metodologia de recolha; convenções de metadados; licenciamento; práticas de armazenamento e segurança; políticas de acesso e partilha; políticas de reutilização e redistribuição; questões éticas e legais (privacidade, consentimento, propriedade intelectual, dados sensíveis); responsabilidades na gestão e curadoria; despesas com curadoria, armazenamento e preservação.

- **Definir as responsabilidades dos/as intervenientes no processo de GDI**. Aconselha-lhe a definição dos papéis (ex. gestor e/ou curador de dados) e das responsabilidades dos/as investigadores/as e dos membros de equipa de investigação no processo de gestão dos dados, identificando os/as responsáveis pela atualização e cumprimento do PGD e fixando as tarefas necessárias, assim como as respetivas autorizações para aceder, organizar, analisar, transformar e divulgar os dados de investigação.

- **Prever os custos associados à GDI**. A previsão das despesas deve ter em linha de conta todas as fases do ciclo de vida dos dados de investigação, tendo em atenção os custos associados com a contratação de recursos humanos, aquisição de material e contratação de serviços especializados. Deve considerar os custos

relacionados com o depósito, preservação e curadoria dos dados de investigação, mediante a realização de uma previsão atempada das necessidades de armazenamento dos dados.

- **Consultar os serviços da NOVA FCSH**. O planeamento das práticas de GDI deve ser feito em articulação com os serviços de apoio à Investigação da NOVA FCSH, nomeadamente a Divisão de Apoio à Investigação (dai@fcsh.unl.pt) e a Divisão de Informática e Transformação Digital, através do Núcleo de Desenvolvimento Digital da Investigação (nddi@fcsh.unl.pt). Deve ser desenvolvido em colaboração com os serviços de consultaria existentes, nomeadamente o responsável pela privacidade de dados (rgpd@fcsh.unl.pt), a Comissão de Ética da NOVA FCSH e o DPO (d*ata protection officer*) da Universidade NOVA de Lisboa (dpo@unl.pt), responsáveis pelo apoio no âmbito da proteção de dados, ética e tratamento de dados sensíveis em CSAH.

Recolha e organização

- **Recolher dados de investigação, acompanhados de metainformação**. Os dados primários e secundários recolhidos durante o processo de investigação devem ser acompanhados de metadados descritivos, seguindo padrões normalizados e reconhecidos internacionalmente (ex. DCMI Metadata Terms), incluindo campos como o título, autor, data e descrição, entre outros. Estes metadados permitem descrever, explicar e localizar os dados de investigação.
    - Sempre que necessário, devem criar-se documentos de consentimento que contemplem as autorizações necessárias para a recolha e a partilha dos dados de investigação.

- **Recolher dados de investigação reutilizados, processados ou modificados**, potenciando a reutilização dos dados de investigação, de acordo com as recomendações do *Annotated Grant Agreement* publicado pela Comissão Europeia em abril de 2023.

- **Organização dos dados de investigação**. Para a organização dos dados é fundamental definir a modalidade de organização dos dados, em ficheiros simples ou bases de dados, e fixar a tipologia, qualidade e formato dos dados, privilegiando formatos abertos e não proprietários (ex. .txt, .odt, .csv, .png, .wav, .mov, .avi, .html, .xml). Importa ainda:

- Organizar logicamente os dados, ficheiros e pastas, assegurando o correto versionamento dos mesmos. O *UK Data Service* disponibiliza um conjunto de boas práticas para auxiliar os/as investigadores/as na estruturação de pastas e ficheiros e na identificação das diferentes versões dos mesmos;
- Assegurar a segurança dos dados, mediante a proteção de ficheiros através de palavras-chave e/ou encriptação do conteúdo.

Processamento e análise

- **Tratar os dados de investigação**. O tratamento e a análise dos dados de investigação permitem criar conjuntos de dados de investigação, conjuntos de dados derivados e *outputs* de natureza distinta, mediante a aplicação de diferentes metodologias, nomeadamente a transcrição de dados, verificação, limpeza, anonimização e/ou tratamento estatístico. Todo o processo relacionado com o processamento dos dados deve ser documentado pelos/as investigadores/as ou pela equipa do projeto, através da criação de notas, documentos de apoio e ficheiros *README* que expliquem o objetivo, metodologia e tipologias documentais resultantes do processamento da informação.

- **Identificar as ferramentas utilizadas no processamento e análise de dados**. Na fase de processamento e análise importa documentar as ferramentas / *software* utilizados no tratamento da informação. Recomenda-se a utilização de *software* livre, flexível e de código-fonte aberto, que permita a organização da informação de forma estruturada.

- **Utilizar dados interligados (*linked data*) nas CSAH**. É indispensável apostar no desenvolvimento dos dados interligados no âmbito das CSAH e generalizar a sua aplicação, pelo potencial de reutilização dos dados que apresenta. O recurso a serviços de vocabulários controlados (ex. ontologias, tesauros), por exemplo, potencia a ligação dos *datasets* criados a conjunto de dados pré-existentes, contribuindo para uma maior abertura e transparência do processo científico.

Publicação e partilha

- **Determinar os direitos de autor sobre os conjuntos de dados**. Normalmente, os/as autores/as dos dados detêm os direitos de autor sobre os mesmos. Em certos casos, os direitos de autor poderão ser detidos pelo

empregador, dependendo dos termos do contrato ou bolsa de investigação. Para mais informações, consultar o "Toolkit for Researchers on Legal Issues" do OpenAire (https://doi.org/10.5281/zenodo.2574619).

- **Determinar o acesso aos conjuntos de dados**. É recomendado que os conjuntos de dados sejam disponibilizados em acesso aberto, segundo o princípio "tão aberto quanto possível, tão fechado quanto necessário". Existem, todavia, exceções que poderão determinar a restrição do acesso aos conjuntos de dados, ou a imposição de um período de embargo. Incluem-se nestas exceções constrangimentos como a exploração comercial dos dados, a privacidade, a confidencialidade e a propriedade intelectual.

- **Selecionar uma licença adequada para partilha dos conjuntos de dados**. Em regra geral, é recomendada a licença Creative Commons CC0 para disponibilização em domínio público de conjuntos de dados de investigação. Para mais informações, consultar o "Toolkit for Researchers on Legal Issues" do OpenAire (https://doi.org/10.5281/zenodo.2574619).

- **Arquivar os conjuntos de dados em repositórios confiáveis**. Na maior parte dos casos, é recomendado o Zenodo (zenodo.org), um repositório generalista muito utilizado para arquivo de conjuntos de dados. Está atualmente em desenvolvimento o Polen (polen.fccn.pt), destinado aos beneficiários de mecanismos de financiamento da FCT. Os portais re3data.org e fairsharing.org incluem informação sobre a generalidade dos repositórios de dados confiáveis.

- **Criar metadados de descoberta dos conjuntos de dados**. A descrição dos conjuntos de dados é fundamental para a sua citação e descoberta através da *Web*. Recomenda-se o uso de esquemas de metadados normalizados e amplamente utilizados na comunidade ou área científica em questão. O arquivo de conjuntos de dados de investigação em repositórios confiáveis requer o preenchimento de um certo número de campos de metadados, normalmente através de esquemas como o DataCite ou o Dublin Core.

- **Considerar a criação de um *data paper***. Um *data paper* é um artigo científico com revisão de pares onde é descrito um conjunto de dados, incluindo a sua recolha, preparação e curadoria. A publicação de um *data paper* é considerada um bom instrumento para valorizar e dar visibilidade aos conjuntos de dados enquanto resultados de investigação.

- **Disseminar os dados de investigação de forma estruturada e dinâmica.** Recorrer a aplicações de código fonte abertas, como sistemas de gestão de conteúdos (ex. Drupal, Omeka), para potenciar a disseminação e a divulgação dos dados de investigação. Estas aplicações permitem trabalhar localmente ou em rede e dispõem de *plugins* que permitem a configuração de protocolos interoperáveis, como o protocolo OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting).

Preservação

- **Migrar para formatos e media adequados para preservação**. Para que o acesso aos dados seja possível a longo prazo, é fundamental que sejam utilizados, sempre que possível, formatos abertos, estáveis, interoperáveis e normalizados para os diversos tipos de media, seja texto, tabelas, imagem, áudio ou vídeo. Formatos proprietários não devem ser utilizados para preservação. Para mais informações, consultar o documento "Data Formats for Preservation" do OpenAire (https://doi.org/10.5281/zenodo.4041512).

- **Armazenar os dados e gerir as versões**. No que diz respeito ao armazenamento dos dados, deve-se distinguir entre dados em bruto e dados processados ou que já foram alvo de análise. Deve ser dada especial atenção ao armazenamento e cópia dos dados em bruto, uma vez que poderão conter informação única e irrepetível (ex. informação de data e hora nos ficheiros). No decorrer da investigação, poderão ser definidas diversas versões dos mesmos conjuntos de dados. Neste caso, deverá ser levada a cabo uma seleção das versões relevantes para fins de preservação, sendo descartadas as restantes. Para mais informações, consultar o documento "Raw Data, Backup and Versioning" do OpenAire (https://zenodo.org/record/4041557).

- **Realizar cópias de segurança (*backups*)**. A realização periódica de cópias de segurança é a melhor forma de prevenir a perda de dados no decorrer das atividades de investigação. Recomenda-se que as cópias de segurança sejam armazenadas em diversos dispositivos, incluindo o armazenamento em nuvem institucional (Microsoft OneDrive), os discos externos e o armazenamento local. As cópias de segurança poderão assumir diversas modalidades:

  - Completas, onde se salvaguardam todos os dados.
  - Diferenciais, onde se salvaguardam todas as alterações desde a última cópia de segurança completa. Neste caso, a recuperação dos dados requer a última cópia completa e a última cópia diferencial.

• Incrementais, onde se salvaguardam todas as alterações desde a última cópia de segurança. Neste caso, a recuperação dos dados requer a última cópia completa e as diversas cópias incrementais que se tiverem realizado.

A nível de software, os sistemas operativos mais utilizados atualmente incluem ferramentas para a realização de cópias de segurança, tais como o Backup and Restore (Windows), o Time Machine (MacOS) e o Déjà Dup (Linux). Para mais informações, consultar o documento "Raw Data, Backup and Versioning" do OpenAire (https://zenodo.org/record/4041557).

Reutilização

- **Encontrar conjuntos de dados**. Existem diversos catálogos e portais agregadores de dados de investigação, de forma a promover a reutilização dos conjuntos de dados. Elencam-se em seguida alguns dos catálogos e portais mais relevantes:

  • Plataforma Aberta de Dados Públicos Portugueses: https://dados.gov.pt/

  • Portal de Dados Europeus: https://data.europa.eu/

  • OpenAire Explore: https://explore.openaire.eu/

  • Google Dataset Search: https://datasetsearch.research.google.com/

- **Verificar termos de acesso, licenciamento e citação dos conjuntos de dados**. Para a reutilização de conjuntos de dados, é fundamental verificar qual a modalidade de acesso através da qual estão disponíveis, tendo em conta possíveis restrições e períodos de embargo. Em paralelo, deverão ser verificados os termos do licenciamento dos conjuntos de dados, que poderão incluir ressalvas, por exemplo, para a exploração comercial, ou menção da autoria dos conjuntos de dados reutilizados. Em qualquer dos casos, encoraja-se a citação dos conjuntos de dados de forma análoga às monografias e aos artigos em revistas científicas, uma vez que se trata de resultados de investigação. Os repositórios de dados frequentemente sugerem uma forma de citação para os conjuntos de dados, a qual deverá ser adaptada às necessidades específicas da investigação. As ferramentas de gestão de referências, como o Zotero (zotero.org), normalmente incluem a tipologia *dataset*, a qual deverá ser utilizada para o caso dos conjuntos de dados.

- **Utilizar conjuntos de dados no ensino e aprendizagem**. Uma boa forma de valorizar os conjuntos de dados é a sua reutilização em atividades de ensino e aprendizagem, por exemplo como base para trabalhos de projeto por parte dos alunos, ou em eventos mais especializados, como oficinas, *masterclasses* ou *hackatons*.

Glossário e abreviaturas

Dados de investigação – Informação coligida, observada, gerada ou criada para validar resultados de investigação originais.
Dados interligados – Paradigma para a publicação de dados estruturados na web e sua interligação com outros dados através de tecnologias e protocolos normalizados.
Esquema de metadados – Definição dos elementos dos dados e das regras no seu uso para a descrição de recursos.
Ficheiro README – Ficheiro onde são documentados os dados de investigação, facilitando o seu entendimento, replicação e reutilização.
FCT – Fundação para a Ciência e Tecnologia.
GDI – Gestão de dados de investigação.
I&D – Investigação e desenvolvimento.
Metadados – Descrição ou conjunto de características de outros dados.
NDDI – Núcleo de Desenvolvimento Digital da Investigação da Divisão de Informática e Transformação Digital da NOVA FCSH.
PGD – Plano de gestão de dados.
Plano de gestão de dados – Documento no qual se especificam todas as questões relacionadas com a recolha, processamento, análise, partilha, preservação e reutilização de conjuntos de dados no âmbito de um projeto de investigação.