RCAAP
Repositório Científico de Acesso Aberto de Portugal

WP4 – D25 – KIT SOBRE DADOS DE INVESTIGAÇÃO

ABRIL DE 2017

GOVERNO DE PORTUGAL | MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA
FCT Fundação para a Ciência e a Tecnologia
Universidade do Minho
Financiado por POS_Conhecimento | União Europeia FEDER

VERSÃO

Autor: José Carvalho; Filipe Furtado; Pedro Príncipe

Contribuição:

Versão: 0.1

Distribuição: [Equipa de Projeto]

Data de Criação: 25 de março de 2014

Última Atualização: 27 de abril de 2017

CONTROLO DE REVISÕES AO DOCUMENTO

Revisão	Data	Alterado por	Alterações

REGISTO DE ALTERAÇÕES

Revisões ao Plano do projeto

GOVERNO DE PORTUGAL | MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA FCT Fundação para a Ciência e a Tecnologia Universidade do Minho

Financiado por POS_Conhecimento Programa Operacional Sociedade do Conhecimento UNIÃO EUROPEIA FEDER

ÍNDICE

INTRODUÇÃO

Este documento define o conteúdo do "Kit de Dados de Investigação" em desenvolvimento no âmbito do projeto RCAAP, tendo como objetivo informar e apoiar, em primeiro lugar, os gestores de repositórios que constituem a comunidade RCAAP e que desenvolvem ações de disseminação e formação nas instituições de investigação e, por consequência, apoiar igualmente investigadores, gestores de ciência e outros parceiros envolvidos em atividades de suporte à gestão de dados de investigação.

Este recurso em língua portuguesa servirá de suporte para futuras iniciativas de gestão de dados de investigação no contexto do RCAAP, podendo ser utilizado de forma direta e independente como informação de apoio, mas também no âmbito dos recursos e-learning do RCAAP, já que o conteúdo aqui definido será devidamente adaptado e disponibilizado na plataforma de e-learning do RCAAP bem como em outros sistemas e recursos de formação do projeto.

Descrição do Kit de Dados de Investigação

O Kit de Dados tem como por objetivo contribuir para colmatar a lacuna que persiste a nível nacional no que respeita informação e materiais disponíveis na área de gestão de dados de investigação. Desta forma, o Kit de Dados pretende dar um contributo significativo para os gestores de repositórios contribuírem para dinamizar uma "cultura de gestão de dados", tanto a nível da comunidade científica como dos profissionais que exerçam a sua atividade no suporte aos investigadores, enquanto gestores de dados ou gestores de ciência.

Esta é uma crescente preocupação no âmbito da comunidade RCAAP, não só mediante a existência da política de dados abertos por parte da FCT[1], mas também devido à obrigatoriedade de cumprimento de semelhantes políticas a nível de outros financiadores, nomeadamente no quadro do programa Horizonte 2020 e dos requisitos a Comissão Europeia para os dados abertos de investigação.

O desenvolvimento e publicação do Kit de Dados beneficia da recente realização dos primeiros eventos[2] a nível nacional exclusivamente dedicados a esta temática: a conferência "Dados de Investigação e Ciência Aberta: rumo a uma estratégia nacional" e do "1º Fórum de Gestão de Dados de Investigação", realizados nos dias 22 e 23 de setembro, respetivamente, na Faculdade de Psicologia e de Ciências de Educação da Universidade do Porto. Nestes

[1] https://www.fct.pt/documentos/PoliticaAcessoAberto_Dados.pdf
[2] http://confdados.rcaap.pt/

eventos, foi reforçada a necessidade de informar e de aproximar a comunidade científica destes temas.

De referir é também o lançamento do portal de Ciência Aberta, recentemente tornado público.[3] Este é um projeto do Ministério da Ciência, Tecnologia e Ensino Superior, que reúne informação, divulga iniciativas e conteúdos formativos nesta área. Tem como público-alvo todos os agentes envolvidos no sistema científico nacional e a sociedade em geral, denominando-se um projeto colaborativo, em desenvolvimento, feito para a comunidade e com a comunidade.

Terminologia de Gestão de Dados de Investigação: Glossário

Considerando que esta é uma área disciplinar e de trabalho recente, é necessária a criação e adoção de uma linguagem comum entre diferentes comunidades dos vários domínios científicos e diferentes *stakeholders*. Neste sentido, salienta-se a existência do glossário[4] recentemente publicado pelo programa Ciência Aberta, que constitui uma primeira referência a nível nacional. No entanto, e sendo notória a ausência de alguns conceitos-chave específicos da gestão de dados, este Kit irá usar como referência para sua terminologia o glossário publicado pelo consórcio CASRAI[5], devido ao seu reconhecimento pela comunidade internacional.

A longo prazo, está prevista a criação de um glossário específico para a gestão de dados de investigação, integrado no Kit e na plataforma de e-learning do RCAAP, procurando a inclusão de tal terminologia específica no glossário já publicado pelo projeto Ciência Aberta.

[3] http://www.ciencia-aberta.pt/
[4] http://www.ciencia-aberta.pt/glossario
[5] dictionary.casrai.org/

KIT DE DADOS DE INVESTIGAÇÃO

1 - Gestão de Dados de Investigação: relevância e desenvolvimentos

Dados de investigação científica constituem a base para todo e qualquer resultado científico e, consequentemente, para toda a publicação ou output científico, baseada em medições, observações ou pesquisas. A gestão de dados é por isso a fundamental para a sua utilização plena e para a validação dos respetivos resultados. Adicionalmente, uma gestão efetiva de dados de investigação permite a possibilidade a publicação independente de pacotes de dados, o que fomenta *i*) a sua reutilização dentro de outros contextos e *ii*) a publicação de dados "negativos".

Adicionalmente, a gestão e publicação de dados de investigação, independente ou adjacente a artigos científicos, assegura a transparência dos mesmos e do processo científico. Tal sucede não só após a publicação dos manuscritos como também durante o processo de revisão por pares.

À semelhança do que acontece com o Acesso Aberto a publicações, o Acesso Aberto a dados de investigação constitui uma forma de democratização do conhecimento científico e de rentabilização do investimento na produção dos mesmos.

2 - Introdução aos Dados de Investigação

Aqui serão introduzidos alguns conceitos chave nesta temática. Esta breve introdução deverá facilitar a compreensão do ciclo de dados de investigação, apresentado na secção seguinte.

Dados de Investigação: o que são

No contexto do RCAAP e deste Kit, consideram-se dados de investigação científica, todos e quaisquer dados de investigação[6] que sejam produto direto ou indireto do processo de investigação científica e por isso necessários para a validação de resultados científicos. A título exemplar e baseado no glossário publicado pelo consórcio CASRAI[7] destacamos

[6] No contexto do RCAAP enquanto repositório digital, não são considerados dados ou documentos em formato analógico.

[7] "Facts, measurements, recordings, records, or observations about the world collected by scientists and others, with a minimum of contextual interpretation. Data may be in any format or medium taking the form of

observações ou registos numéricos, textuais, imagens e vídeos, nos mais variados formatos digitais, enquanto exemplos de dados de investigação. Assim, podem distinguir-se dois tipos diferentes de dados de acordo com o seu grau de processamento, primários e secundários, conforme abaixo detalhado:

i) primários: dados de investigação obtidos diretamente do processo de investigação, instrumento ou metodologia científica, sem que tenham sofrido qualquer processamento ou transformação (p. ex.: entrevista áudio/vídeo sem edição, dados de gerados por um instrumento de medição sem que tenham sofrido processamento).

ii) secundários: dados resultantes da interpretação, processamento ou transformação de dados primários (p. ex.: entrevista áudio/vídeo após edição, dados de gerados por um instrumento de medição sem após processamento ou aplicação de modelos estatísticos).

Dados de investigação fazem portanto parte integrante de qualquer processo de investigação científica, e como tal, são base para todo o "output científico". Em seguida, é analisado em detalhe o ciclo de dados de investigação, de forma a ilustrar a integração e gestão dos mesmos, no processo de investigação científica.

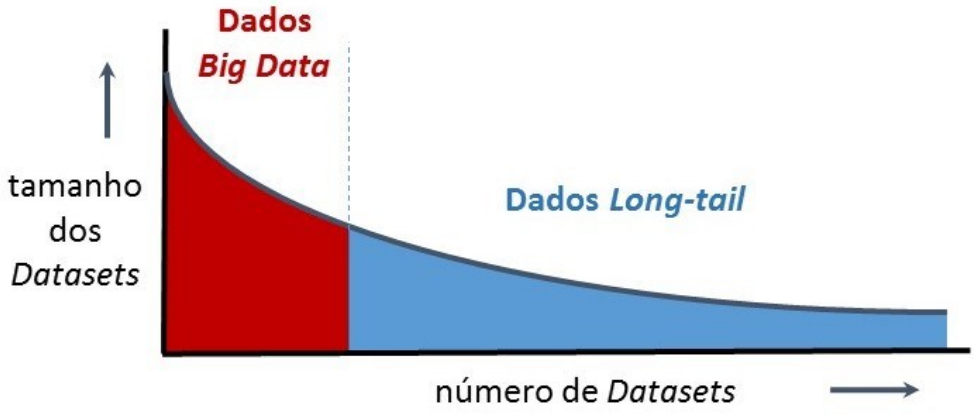
É importante notar que, ainda possa ser útil, a distinção entre dados primários ou secundários não é livre de ambiguidades: diferentes disciplinas poderão ter noções diferentes acerca de um mesmo conjunto de dados.

Big-Data vs. dados "long-tail"

Para além da distinção relativa ao grau de processamento (dados primários ou secundários) é também relevante distinguir diferentes tipos de dados de investigação, nomeadamente clarificar os conceitos "dados *big-data"* e "dados *long-tail"*, uma vez que as suas características implicam diferentes medidas para a gestão de dados.

writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records."; *in*: http://dictionary.casrai.org/Data

A figura seguinte ilustra esquematicamente a distribuição do tamanho de *datasets* produzidos em investigação científica, pelo respetivo número de *datasets*. Existem várias fontes,[8,9] que indicam esta como sendo a distribuição característica da produção atual de dados de investigação, ou seja: a minoria dos conjuntos de dados (*datasets*), corresponde ao maior volume de informação (*big-data*, superfície a vermelho); Inversamente, a larga maioria dos dados de investigação (*datasets*) corresponde à "cauda-longa" da distribuição (*long-tail*, superfície a azul).



A tabela seguinte procura comparar as características destes tipos de dados de investigação, salientando os principais aspetos de relevância para a sua gestão e preservação.

Tipo	***Big-Data***	**Dados *Long-Tail***
Homogeneidade	homogéneos	heterógenos
Volume dos *Datasets*	grande	pequeno
Diretrizes e normas	estabelecidas	únicas ou inexistentes
Curadoria	centralizada	individual
Tipo de Repositórios	disciplinar	institucional
Reutilização dos Dados	frequente	rara

Tabela 1: Tabela-resumo das características de dados de investigação big-data *e dados* long-tail.[10]

[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728080/; Nat Neurosci. 2014 Nov; 17(11): pp. 1442–1447; doi: 10.1038/nn.3838

[9] http://science.sciencemag.org/content/331/6018/692; *Science* 11 Feb 2011: Vol. 331, Issue 6018, pp. 692-693; DOI: 10.1126/science.331.6018.692

[10] Adaptado do artigo *Shedding Light on the Dark Data in the Long Tail of Science,* P. Bryan Heidorn. 2008

A implicação prática destes conceitos e de tal distribuição é o facto da maioria dos dados de investigação – Dados "*long-tail*" - constituírem o maior desafio em termos de planeamento, gestão, preservação e reutilização, devido à sua natureza heterogénea e singular.

Preservação Digital

Por preservação digital entende-se, neste contexto, um conjunto de medidas e infraestruturas necessárias para garantir a integridade de dados digitais a longo-termo, por vários anos e idealmente por tempo o termo indeterminado (*ad eternum*). Inclui estratégias de backups regulares, algoritmos de sincronização, migração para outros formatos físicos, entre outras medidas. A título complementar inclui-se aqui também a definição do programa Ciência Aberta[3], que define "preservação"[11] como "um termo genérico que designa o conjunto de medidas a empreender para garantir a preservação da integridade dos documentos e dos seus conteúdos, em manter o acesso eletrónico e em assegurar a legibilidade e a perenidade dos seus conteúdos durante um longo período de tempo".

3 - Ciclos de Gestão de Dados de Investigação

Um ciclo de gestão de dados de investigação procura integrar os conceitos e os momentos-chave no planeamento e implementação de ações relacionadas com a gestão de dados de investigação. Esses conceitos e ações variam necessariamente consoante a disciplina e o tipo de dados produzidos, entre outros fatores. Por esta razão, são apresentados em seguida três diferentes ciclos de dados e de investigação, de diferentes fontes. Por último, é apresentado um quarto ciclo de dados de investigação, que procura resumir e combinar os vários conceitos apresentados, bem como uma discussão detalhada dos mesmos.

[11] http://openaccess.inist.fr/spip.php?page=glossaire

Ciclos de Dados de Investigação (fontes externas)

a) Universidade de Queensland

A Universidade de Queensland apresenta uma visão generalista e global do processo de gestão de dados, com o ciclo de vida de dados de investigação acima apresentado,[12] focando-se nas fases de recolha, processamento, análise, preservação, descoberta e reutilização.

[12] http://guides.library.uq.edu.au/research-data-management

b) Universidade da Califórnia

A Universidade da Califórnia[13] apresenta um ciclo de Investigação Científica, onde é focado o planeamento, a implementação, disseminação, descoberta, preservação e, finalmente, a reutilização do output produzido.

[13] http://www.lib.uci.edu/dss

c) Anthony Beitz

Este esquema criado por Anthony Beitz[14] fornece a visão a mais simplista dos esquemas aqui apresentados, focando as diferentes fases de conceção e planeamento (a branco e amarelo claro), a fase de execução na plataforma para a gestão de dados e, finalmente, a fase de disseminação.

[14] Slides de uma apresentação por Anthony Beitz na conferência "Open Repositories 2012"

Ciclo de Dados de Investigação: RCAAP

O ciclo de dados seguinte procura não só resumir os previamente apresentados, como também combinar três fases distintas, no processo de gestão de dados de investigação: Fases de Planeamento, Produção e Disseminação.

Fase de Planeamento

Na Fase de Planeamento dever-se-ão fazer as primeiras reflexões quanto à produção, preservação e partilha de dados de investigação, idealmente formalizadas num documento para o efeito, denominado de Plano de Gestão de Dados (PGD). Frequentemente, a submissão de um PGD constitui um requisito de financiadores de ciência, aquando da submissão de projeto e concurso a financiamento.

Fase de Produção

Após o início do Projeto e dos trabalhos de investigação são criados os primeiros dados no âmbito desse mesmo projeto. Nesta fase ocorrem todos os procedimentos e transformações aos dados, para que possam ser posteriormente publicados, passando deste modo do domínio restrito ao domínio público.

Fase de Disseminação

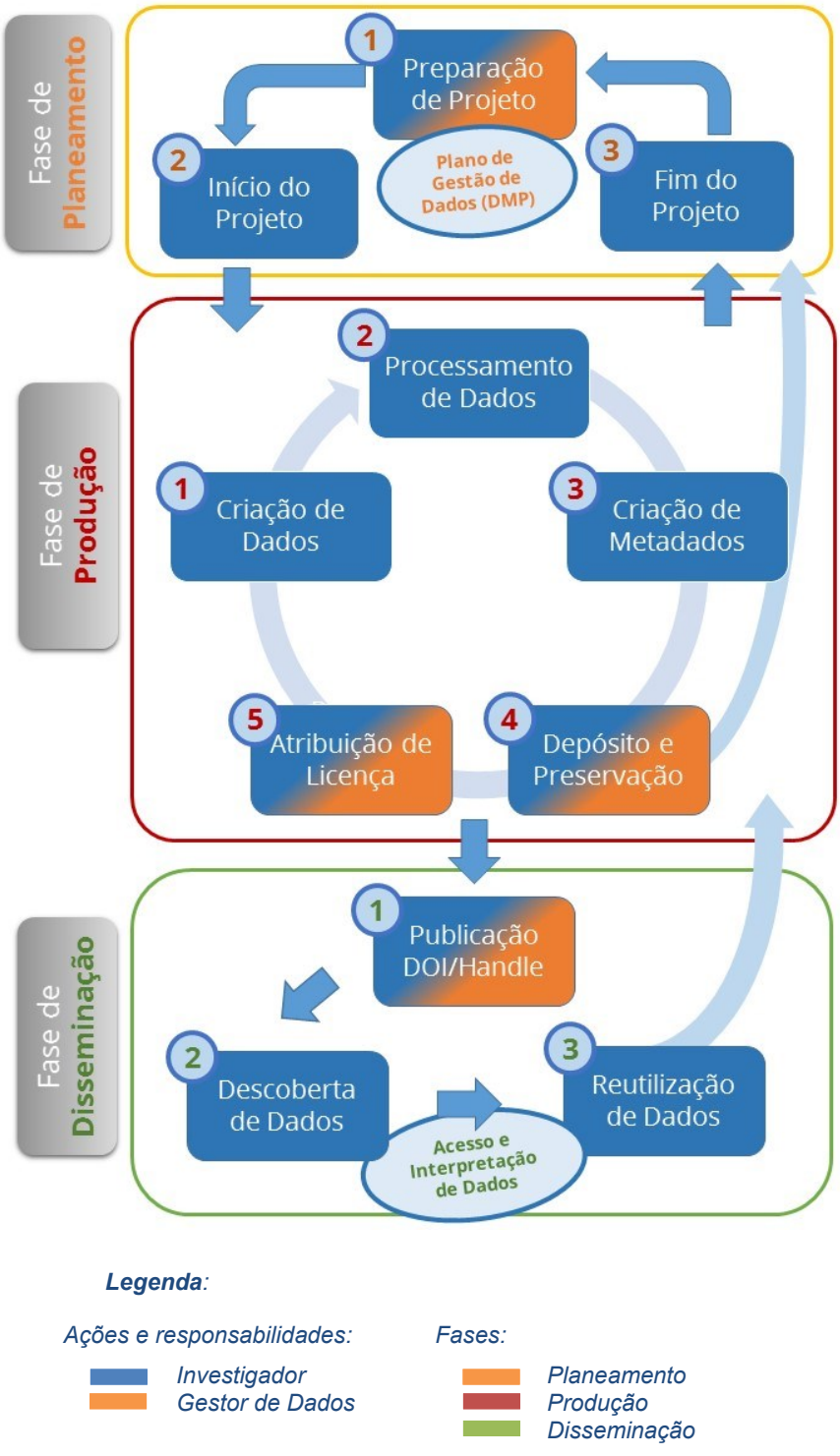
É após publicação dos dados que estes podem finalmente ser acedidos e reutilizados, gerando eventualmente, novos dados de investigação, e dando novamente reinício ao ciclo de dados.

A reutilização de dados é o objetivo final e central da implementação de estratégias de gestão e preservação de dados, constituindo o colmatar do ciclo e do processo que gera em si, o valor acrescentado aos dados de investigação produzidos (no domínio restrito).

É importante notar que podem existir casos, em que os dados produzidos são imediatamente publicados, ou seja, de modo em que o momento da produção seja coincidente com o momento da disseminação. No entanto, tal só deverá acontecer, após a reflexão cuidada dos passos apresentados nas fases distintas, pelo que não é considerada boa prática a disseminação sem que tenham existido previamente medidas concretas de curadoria e gestão. Tal não exclui no entanto, a hipótese de estabelecer um automatismo no processo; Tipicamente, estes correspondem a casos de produção de dados estruturados, homogéneos, devidamente descritos e em grande escala, o que corresponde a dados *big-data*.

Diagrama do ciclo de gestão de dados de investigação RCAAP

A figura seguinte procura ilustrar em detalhe as três fases acima referidas e a interligação dos diferentes conceitos-chave. Esta ilustração representa um *workflow* genérico e, por isso, deve notar-se que diferentes projetos de curadoria ou repositórios de dados poderão ter *workflows* específicos e por isso diferentes do abaixo ilustrado.



Fase de Planeamento

1. Preparação de Projeto

No contexto da gestão de dados de investigação, a preparação do projeto requer uma reflexão cuidada acerca dos tipos de dados, da infraestrutura e de *workflows* para a partilha dos mesmos. Idealmente, estas considerações serão idealmente formalizadas num PGD[15].

Poderá ser necessária uma estimativa do volume de a serem criados, bem como soluções concretas para o armazenamento e preservação dos mesmos, bem como estimativas dos custos associados; Para tal refere-se aqui o projeto 4C (Collaboration to Clarify the Cost of Curation),[16] que fornece informação nesta tarefa.

Em caso de submissão de candidatura para financiamento do projeto, o financiador pode requerer dados de investigação de projetos anteriores, nomeadamente, indicando onde estes se encontram depositados ou até mesmo requisitar o acesso aos mesmos.

É nesta fase que deverá ser formalizado o plano de gestão de dados, frequentemente, um requisito de financiadores. Da mesma forma, é frequente que o financiador exija alterações durante o decorrer dos trabalhos (Fase de Produção) e no Fim do Projeto (ver abaixo). No caso de projetos financiados pela Comissão Europeia (ver Piloto de Dados Abertos H2020),[15] é necessária a submissão de um plano de gestão de dados durante os primeiros 6 meses de projeto. É este é o tema da secção seguinte, onde é fornecido material para a criação de um PGD.

2. Início do Projeto

Mediante o começo do projeto dá-se início ao processo e às tarefas relacionadas com a investigação científica propriamente dita, entrando-se no ciclo de dados de investigação onde serão gerados os primeiros dados digitais, em domínio restrito.

3. Fim do Projeto

O final do projeto consiste não só em documentar e publicar os resultados obtidos durante o mesmo como também em preparar os próximos trabalhos de investigação e, possivelmente, a submissão de candidaturas para financiamento. Como já referido na fase de Preparação do Projeto, a requisição de dados de projetos anteriores ou a indicação de onde estes se encontram depositados é uma exigência cada vez mais frequente por parte das agências financiadoras, por isso é fundamental ter em conta o próximo projeto, na fase de conclusão.

[15] https://www.openaire.eu/opendatapilot-dmp; https://www.openaire.eu/opendatapilot-dmp
[16] http://4cproject.eu/

Também por esta razão, torna-se fundamental arquivar digitalmente os dados que não foram publicados e assim garantir a preservação a longo prazo dos mesmos.

Fase de Produção

1. Criação de Dados

Como referido na Introdução, consideram-se dados de investigação todos e quaisquer dados digitais que sejam produto direto ou indireto do processo de investigação científica. Exemplarmente destacam-se observações, registos numéricos, textuais, imagens ou vídeos, nos mais variados formatos digitais, enquanto exemplos de dados de investigação.
Deste modo, esta fase do ciclo de dados inicia-se com a produção, criação ou recolha de dados de investigação. Os passos subsequentes são necessários para garantir a (re)utilização sustentável dos dados produzidos.

2. Processamento de Dados

De modo a extrair a informação dos dados produzidos (primários) é frequente a edição, alteração ou seleção dos dados produzidos. São assim criados dados secundários que podem ser posteriormente usados como base para publicações científicas. No entanto, para serem publicados por si só, são necessários vários passos subsequentes, começando pela descrição dos dados obtidos.

3. Criação de Metadados

De um modo geral, os dados produzidos são descritos de forma mais ou menos completa, no momento da sua criação. No entanto, de modo a que os dados produzidos (e processados) possam ser reutilizados pela comunidade científica mas também por cidadãos fora do contexto institucional, é fundamental que a informação descritiva seja tão exaustiva quanto possível. Adicionalmente, é de extrema importância que esta informação seja fornecida da forma mais estandardizada possível, de modo permitir a indexação em bases e portais de dados, procedimentos de "*harvesting*" ou "*data-mining*" ou simplesmente de modo a poder ser encontrada por motores de busca.
O cumprimento destas duas premissas – o mais exaustivo e, simultaneamente, o mais estandardizado quanto possível – constituem um dos desafios na descrição dos dados, por parte da comunidade científica. Aconselha-se por isso a adoção antecipada de um esquema de metadados – idealmente, aquando da elaboração do Plano de Gestão de Dados – e que a descrição dos dados produzidos comece tão cedo quanto possível após a sua produção.

Existem variadas diretrizes, normas e esquemas de metadados, específicos de determinadas disciplinas, ou de determinados repositórios de dados. O esquema de metadados DataCite[17] é um bom ponto de partida, uma vez que foi concebido especificamente e de forma interdisciplinar para dados de investigação. Alguns exemplos dignos de nota são referidos na secção de Identificação de Recursos (ver Metadados).

Mediante a descrição dos dados de investigação, através da criação de metadados é possível a sua agregação, descoberta, acesso, e reutilização; sem descrição os dados criados são impossíveis de ser utilizados por outros que não o seu criador.

4. Depósito e Preservação

O depósito num repositório digital tem a função particularmente importante de assegurar a integridade digital dos pacotes de dados a longo prazo, através de variadas medidas de preservação digital. É portanto um passo essencial para o acesso livre e sustentável a longo prazo. Para além disso, a escolha do repositório está relacionada com o esquema de metadados adotado, de forma a garantir que a descrição dos dados corresponda às diretrizes usadas pelo repositório escolhido ou que seja interoperável com o mesmo.

No que diz respeito à escolha de um repositório de dados, o projeto re3data.org poderá servir como um bom ponto de partida na procura de um repositório adequado à área disciplinar pretendida. Recomenda-se explicitamente a preferência por repositórios de dados certificados, atendendo às diretrizes do respetivo selo de certificação. Existem diferentes agências que certificam repositórios e, por isso, diferentes critérios para a certificação.

No campo da certificação de repositórios, é ainda importante mencionar o grupo de trabalho da Research Data Alliance[18] e as atividades desenvolvidas.

5. Atribuição de Licença

A atribuição de uma licença é um passo fundamental para a partilha de dados que regula o acesso aos mesmos e, por isso mesmo, deve anteceder a sua publicação. Deverá proteger a propriedade intelectual por parte dos detentores - autor(es) e/ou instituição), sem no entanto limitar desnecessariamente a sua reutilização por terceiros.

Apesar de se tratar de um assunto de elevada complexidade, nomeadamente no que respeita a dados de investigação, existem alguns modelos de licenças amplamente utilizados, o que

[17] Datacite metadata: https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf
[18] https://www.rd-alliance.org/groups/repository-audit-and-certification-dsa–wds-partnership-wg.html

facilita muito a sua escolha por parte dos investigadores. Desta forma destacam-se aqui as licenças Creative Commons.[19]

É fundamental que a escolha da licença seja efetuada pelo investigador e/ou detentor dos direitos de autor, em concordância com outros detentores desses direitos. O gestor de dados da instituição pertencente deverá acompanhar este processo e fornecer esclarecimento sempre que necessário.

Mais informação é fornecida neste Kit, na secção "Direitos de Autor e Licenciamento de Dados".

Fase de Disseminação

1. Publicação (DOI/Handle)

Cumprido o ciclo descrito na fase de produção, ou seja, uma vez que os dados tenham sido criados, processados, descritos, preservados e devidamente licenciados, estar-se-á, à partida, em condições de proceder à sua publicação.

Com a publicação dos dados dá-se início à disseminação dos mesmos, ou seja, os respetivos metadados serão passiveis de serem agregados e por isso encontrados e acedidos. Dependendo do tipo, condições de publicação e eventual existência de períodos de embargo, os próprios dados poderão estar também imediatamente acessíveis. A publicação constitui portanto uma condição necessária para a reutilização dos dados, por parte de elementos exteriores ao grupo de trabalho onde estes foram criados.

Particularmente relevante no momento da publicação é a atribuição de um identificador único e persistente ao pacote de dados, de forma a garantir a citação dos mesmos, dos respetivos autores, metadados e outros recursos a ele associados.

Existem variados identificadores persistentes,[20] sendo o Handle e DOI os mais comuns no contexto da gestão de dados.[21]

2. Descoberta de dados de investigação

É importante considerar os aspetos que condicionem ou potenciem a visibilidade de dados publicados, de forma a maximizar a sua descoberta. Em termos gerais, a visibilidade dos dados dependerá do repositório escolhido para o depósito, em particular, dos portais de dados que agregam metadados do mesmo. Adicionalmente, é importante sublinhar novamente, que

[19] http://creativecommons.pt/

[20] Archival Resource Keys (ARKs), Digital Object Identifiers (DOIs), the Handle System, Persistent Uniform Resource Locators (PURLs), Uniform Resource Names (URNs), e Extensible Resource Identifiers (XRIs).

[21] https://hdl.handle.net/; https://dx.doi.org/

quanto mais completo for o esquema e preenchimento de metadados maior será a visibilidade dos dados. Os resultados de uma pesquisa serão certamente também dependentes da ferramenta escolhida para procurar dados de investigação, da qual irá resultar mais ou melhores resultados, dependendo da eficiência do portal agregador ou do motor de busca.

Nota adicional: Acesso, interpretação e descodificação (digital) dos dados

Como descrito até agora, a reutilização de dados de investigação pressupõe a sua descoberta (uma vez no domínio público), o seu acesso (aberto ou não) e condições favoráveis à sua reutilização (regulada pela licença em vigor).

No entanto, para aceder à informação propriamente dita e contida nos pacotes de dados, é fundamental que estes possam ser digitalmente interpretados e descodificados. Para tal, é necessário o uso de software e hardware apropriados. A escolha do formato digital é portanto de grande relevância, devendo-se optar sempre que possível por formatos de ficheiros e dados abertos e não por tipos de ficheiro que pressuponham software proprietário (ao invés de software open-source), de modo a não condicionar a descodificação digital dos mesmos.

3. Reutilização de Dados

Após a descodificação dos ficheiros contidos nos pacotes de dados poder-se-á, à partida, proceder à sua (re)utilização. Desta forma, cumpre-se o objetivo principal da gestão e preservação de dados de investigação.

Existem várias formas em como dados de investigação podem ser reutilizados:

- os dados (primários) podem ser usados para gerar novos dados (secundários): neste caso, mediante a aplicação de um processamento alternativo, seleção e combinação com outros dados, são criados novos dados de investigação, dando-se o reinício do ciclo de dados em domínio restrito.
- os dados podem ser diretamente citados para suportar literatura ou estudos científicos, nomeadamente teses ou artigos: neste caso, não há criação de novos dados, sendo no entanto necessária a citação dos mesmos, para suporte de literatura científica.

Algumas editoras, suportam ainda a possibilidade de integrar os dados na publicação científica em si, de forma interativa ("*enhanced-publication*"),[22] sendo esta uma forma relativamente recente de publicar trabalho científico.

Em qualquer dos casos, a reutilização pressupõe a citação do conjunto original dos dados, fazendo uso do respetivo identificador persistente. É importante ter em conta convenções para a citação de dados, porventura existentes para o contexto em que os dados são reutilizados ou citados. Mais informação na secção de Identificação de Recursos (*xi.*)

[22] https://en.wikipedia.org/wiki/Enhanced_publication

3 - Planos de Gestão de Dados

Esta secção é baseada no documento produzido pelo DCC/JISC "Como Desenvolver um Plano de Gestão de Dados"[23]; Tem como principal objetivo fornecer informação prática acerca de como escrever um PGD para o seu projeto.

Secção 1 - Introdução

Porquê desenvolver um plano de gestão de dados?

- Permite encontrar e entender os seus dados, quando precisa deles
- Reforça a continuidade do projeto, mediante saída ou entrada de pessoal
- Evita a duplicação de dados, na criação/recolha ou processamento dos dados
- Preserva os dados que estão na base das publicações, permitindo a validação posterior dos resultados e conclusões obtidas
- Contribui para um aumento da visibilidade de resultados científicos e da cooperação científica
- Contribui a citação e reutilização dos dados por parte de outros investigadores

O planeamento ajuda a concretizar os objetivos acima mencionados: em última análise, o principal benificiário é o próprio investigador!

A elaboração de um PGD contribui para um trabalho mais efetivo e eficiente, na medida em que facilita o processo de investigação. O planeamento permite que sejam tomadas decisões consequentes, permitindo a escalabilidade das mesmas e a adaptação a imprevistos.

Para além das vantagens referidas, convém relembrar que os financiadores de ciência, editores e revisores "*peer*" podem, em determinadas circunstâncias, eles próprios exigir a partilha de dados de investigação.

O que querem os financiadores de Ciência?

Várias entidades financiadoras de ciência – incluindo a Comissão Europeia - já adotaram políticas de exigir PGD aquando da submissão de projetos, ou até mesmo dados de investigação de projetos anteriores.

Estas políticas podem e devem também ser vistas como uma oportunidade para os cientistas e grupos de investigação de demonstrarem a consciência do que são boas práticas de

[23] http://www.dcc.ac.uk/resources/how-guides/develop-data-plan

investigação científica. Tais práticas que constituem uma forte mais-valia na avaliação do projeto e reassegura o financiador de um bom investimento no projeto proposto.

A ferramenta: DMP-Online

O DCC criou o DMP Online, uma ferramenta para auxiliar os investigadores na criação de um PGD, de acordo com os requisitos específicos de financiadores do Reino Unido. No entanto, a estrutura pode facilmente ser adaptada a outros requisitos, uma vez que a sua estrutura central é baseada numa "*checklist*" de tópicos a abordar. Estes são:

1. Tipos de Dados, Formatos, Normas e Métodos de Captura de Dados
2. Ética e Propriedade Intelectual
3. Acesso, partilha e reuso de dados
4. Armazenamento a Curto Prazo e Gestão de Dados
5. Depósito e Preservação a Longo Prazo
6. Recursos

Considerações acerca de como elaborar o plano

- Consultar e colaborar

É importante considerar as diferentes opções e procurar aconselhamento, de modo a ter em conta o contexto do processo de decisão. É particularmente útil procurar aconselhamento em questões mais técnicas, uma vez que estes afetam a forma como o projeto está planeado, nomeadamente, qual o "*know-how*" necessário para os métodos utilizados relativamente à aquisição, análise e preservação de dados. Dirija-se em particular a colegas, na sua Biblioteca, serviços de IT ou Gestores de Ciência na sua instituição, abordando estes temas.

- Usar o suporte e a experiência existentes

Por vezes, não é necessário "reinventar a roda" e outros já percorreram esse caminho, na própria instituição. Aproveite a experiência de outros, e use modelos para PGD pré-estabelecidos, caso a sua instituição os disponha.

Também é possível obter suporte mais generalista, através de determinados centros ou repositórios de dados.

- Justifique suas decisões

Tipicamente, os financiadores não especificam formatos de arquivo, diretrizes ou metodologias que deve usar na produção dos dados de investigação. Por isso, acima de tudo, é necessário que no PGD as suas opções a este respeito sejam justificadas e contextualizadas, de acordo com vários fatores.

- Esteja preparado para implementar seu plano

Os financiadores irão querer comprovar que os seus requisitos são entendidos e que existem, de facto, planos realistas para os seguir. Para além disso, o PGD é um documento vivo, o que quer dizer que irão ser submetidas diferentes versões do mesmo, contemplando o que foi implementado, o que não foi e porquê.

Secção 2 - Conteúdo de um plano de gestão de dados

Tipos de Dados, Formatos, Convenções e Métodos

É importante justificar as escolhas: deverá ser fornecida informação acerca de qual o tipo de dados que serão criados no âmbito do projeto, e quais os formatos, diretrizes e metodologias utilizadas. Convém não esquecer, que as escolhas efetuadas poderão facilitar ou dificultar a partilha e reutilização de dados. Pode ser útil gerar/recolher dados num formato conhecido pela comunidade de trabalho, tornando os dados interoperáveis. Idealmente, os dados serão partilhados num formato aberto, o que claramente facilita o seu processamento posterior. Convém no entanto respeitar as convenções e diretrizes do repositório escolhido, caso se apliquem.

Ética e Propriedade Intelectual

Apresente argumentos sólidos respeitantes a restrições na partilha de dados. Explique as restrições, tais como, períodos de embargo ou outras restrições, de modo a garantir a justificação clara dos mesmos, uma vez que é frequentemente esperado que a ciência financiada por fundos públicos seja disposta livremente, ao público.

Toda a investigação envolvendo dados ou materiais humanos ou pessoais deverá ser revista, do ponto de vista ético. Em certos casos, poderão ser necessárias medidas de anonimização de dados, de modo a proteger a identidade e informações acerca dos participantes. Estas estratégias deverão ser claramente delineadas e definidas no PGD.

A Propriedade Intelectual dos dados deverá ser clarificada e, quando necessário, planos concretos para a negociação de licenças, no início do projeto de investigação. No caso de haver a decisão de compra ou negociação dos direitos de autor, é importante ter em conta que eventuais restrições irão influenciar diretamente o depósito e a partilha de dados.

Acesso, Partilha e Reutilização de Dados

Planear e antecipar a reutilização de dados: É importante ponderar qual poderá ser o público interessado nos dados que serão produzidos no âmbito do projeto, de modo a otimizar ao modo de partilha dos dados. Adicionalmente, deverão ser considerados os requisitos e diretrizes, específicos dos repositórios de dados, de modo a garantir a qualidade e interoperabilidade dos mesmos.

Especificar os detalhes para o acesso: reassegure os financiadores acerca de onde, quando e como irá disponibilizar os seus dados; Financiadores de ciência esperam frequentemente a disponibilização dos respetivos prazos. No caso do não cumprimento dos prazos estabelecidos, é importante uma justificação fundada, especificando quais os obstáculos, dificuldades e restrições, bem como qual a estratégia delineada para os ultrapassar.

Usar uma infraestrutura existente: sempre que possível, escolha um repositório estabelecido, centro de dados, ou repositório institucional. Caso esteja incerto de quais os serviços à sua disposição, consulte a lista de repositórios no projetos DataCite, BiomedCentral, e no Re3data. Caso, o acesso aos dados tenha de ser restrito, procure serviços de dados seguros ou arquivos de dados.

Gestão e Armazenamento de Dados a Curto Prazo

Defina o apoio à gestão de dados: É fundamental identificar quais os recursos que se encontram disponíveis e quais os que precisam de ser desenvolvidos. No caso de haver suporte local disponível, é relevante demonstrar que existe um trabalho e uma ajuda mútua, de modo a poder usufruir da mesma. No caso de ser necessário suporte técnico ou outros serviços de um parceiro externo justifique as decisões tomadas e o orçamento respetivo.

Depósito e Preservação a Longo-Prazo

Selecione dados que sejam valiosos a longo-prazo: Partilha e preservação de dados poderá não ser justificável em todos os casos. É por isso importante separar e selecionar quais os dados que trazem realmente uma mais-valia e que cumprem esta finalidade.

Guarde os dados que suportam textos- ou figuras-chave do seu trabalho: Particularmente relevante são os dados que suportam a informação científica, seja ela gráfica ou textual.

Definir Recursos, Estimar Custos

Defina e justifique recursos: Procure estimar e especificar os recursos necessários, de modo a poder apresentar um orçamento realista. O orçamento deverá discriminar e justificar os custos relacionados com a gestão de dados de investigação, sem no entanto subestimar os custos e o empenho de humanos.

Secção 3 - Passos práticos para começar

Exemplos: Alguns projetos dispõem publicamente o seu PGD, podendo este ser livremente consultado. Deste modo, poderá observar alguns documentos que, a título exemplar, poderão dar uma ideia de como formular o seu próprio PGD. Encontrará vários exemplos na página do DCC[24]. Consulte também a Secção 5 - Políticas e Diretrizes de Dados de Investigação: Politicas Institucionais, deste Kit.

4 - Direitos de Autor, Licenciamento de Dados e Dados Pessoais

Esta secção procura salientar algumas questões fundamentais no contexto da atribuição de licenças à comunidade científica, e de fomentar um sentido crítico neste campo. Deverá, por isso, ser encarada e usada como um ponto de partida nesta temática, não sendo um recurso exaustivo e final. Como informação adicional aos conteúdos apresentados nesta secção destacamos o workshop realizado no âmbito do 2ºFórum-GDI, "Questões legais: proteção, licenciamento e reutilização de dados".[25]

Licenças Creative Commons

As licenças Creative Commons (CC) são um conjunto de licenças amplamente usadas no licenciamento de trabalhos científicos, nos quais se inserem os dados de investigação e bases de dados. Isto deve-se, essencialmente, à simplificação que introduziram na comunidade científica, num assunto que é por natureza bastante complexo, permitindo a atribuição de licenças a trabalhos científicos de uma forma estandardizada e reconhecida internacionalmente.

Apesar de amplamente usadas, existem algumas considerações fundamentais que os detentores dos direitos de autor devem fazer, ao atribuir licenças Creative Commons (CC).

[24] http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples

[25] Teresa Nobre, Advogada e Coordenadora Jurídica da Creative Commons Portugal:
Apresentação: http://forumgdi.rcaap.pt/wp-content/uploads/2017/04/ForumDados_TeN_310317.pdf;
Vídeo: https://educast.fccn.pt/vod/clips/24wxse4jl6/flash.html?start=0:00:00:00.000&end=0:02:07:15.822

Considerações na atribuição de Licenças

Direitos de Autor

Em primeiro lugar, é sempre aconselhável a verificação da detenção dos direitos de autor, uma vez que nem sempre o autor é detentor (exclusivo) dos mesmos. Tais considerações são válidas em casos de autoria partilhada e tornam-se particularmente relevantes para trabalhos produzidos em cooperação entre vários institutos, por vezes a nível internacional. Nestes casos, os vários institutos ou autores podem ter diferentes países de origem e, eventualmente, diferentes formas de regular o acesso a obras científicas.

Limite de Originalidade

É também importante considerar, que dados de investigação, por si só, não são passiveis de serem protegidos por licenças ou direitos de autor: é fundamental que os mesmos excedam o chamado "limite de originalidade". Ou seja, para a atribuição de uma licença é necessário que este limite seja excedido sendo, nesse caso, o conjunto de dados designando por "obra original".

Estes princípios aplicam-se também à proteção a bases de dados, sendo a atribuição de uma licença possível mediante o cumprimento do requisito do limite de originalidade ou a demonstração de um esforço considerável para a sua elaboração. Neste contexto, recomenda-se vivamente a leitura da publicação OpenAIRE "Safe to be Open: Study on the protection of research data and recommendation for access and usage".[26] Encontra-se também disponível *online* um resumo e as principais conclusões deste trabalho.[27]

Irrevogabilidade das Licenças CC

Adicionalmente, consideramos fundamental salientar a irrevogabilidade das Licenças CC. Em termos práticos, isto implica que após o acesso a um trabalho ou a conjunto de dados com uma determinada licença, o benificiário terá permanentemente o direito à sua reutilização, nos termos dessa licença, ainda que esta seja posteriormente alterada.[28] Ou seja, este acesso é

[26] "Safe to be open: Study on the protection of research data and recommendations for access and usage", Universitätsverlag Göttingen; Guibault, Lucie; Wiebe, Andreas (Eds) (2013); ISBN: 978-3-86395-147-4.
[27] https://www.openaire.eu/public-documents?id=789&task=document.viewdoc
[28] "The CC licenses are irrevocable. This means that once you receive material under a CC license, you will always have the right to use it under those license terms, even if the licensor changes his or her mind and stops distributing under the CC license terms. Of course, you may choose to respect the licensor's wishes and stop using the work." *in* https://creativecommons.org/faq/

irrevogável, mesmo que o autor decida posteriormente retirar o trabalho ou atribuir-lhe uma licença mais restritiva.

Lei Portuguesa e Internacional

O documento que regula e define conceitos legais relacionados com esta temática – o "Código do Direito de Autor e dos Direitos Conexos" - encontra-se disponível.[29]

No contexto da União Europeia e em Portugal, os direitos de autor seguem o modelo francês (*droit d'auteur*, ou *Urheberrecht*, na Alemanha) por oposição ao conceito de Copyright, o modelo estabelecido nos Estados Unidos. É importante notar que existem diferenças fundamentais entre estes modelos, como discutido no artigo *online*.[30]

No que respeita o modelo europeu, aconselha-se a consulta do material de apoio disponibilizado pelo DCC[31] (Reino Unido) bem como da informação disponível no site Open Definition,[32] em particular a secção acerca de dados de investigação.

O Institute for Information Law (IViR)[33] da Universidade de Amsterdão publicou um artigo[34] exaustivo acerca desta temática, cuja leitura é aqui explicitamente aconselhada.

Uma outra referência que lida especificamente com a lei de copyright no Reino Unido e aplicabilidade a sistemas de informação e tecnologia de informação é publicada pelo Jisc.[35]

Referimos, novamente, o trabalho publicado pelo projeto OpenAIRE.[26]

Dados de Investigação e Dados Pessoais

A recolha, processamento e divulgação de dados pessoais é um tema sensível que deverá ser considerado e planeado desde o início dos trabalhos de investigação ou início do respetivo projeto. Neste contexto, é fundamental garantir o respeito à lei e a constituição portuguesas, salientando-se de seguida trechos de algumas leis e artigos, não dispensando no entanto a leitura integral dos mesmos por parte de autores, investigadores ou detentores dos direitos de autor:

- a Lei n.º 67/98 (artigo 3º), publicada em Diário da República[36] que define "Dados pessoais" como "qualquer informação, de qualquer natureza e independentemente do respetivo suporte,

[29] https://www.spautores.pt/assets_live/165/codigododireitodeautorcdadclei162008.pdf
[30] http://www.sacd.fr/Author-s-rights-vs-copyright.2146.0.html
[31] http://www.dcc.ac.uk/resources/how-guides/license-research-data
[32] http://opendefinition.org/guide/data/
[33] http://www.ivir.nl/
[34] http://www.ivir.nl/publicaties/download/Open_Research_Data.pdf
[35] https://www.jisc.ac.uk/guides/copyright-law
[36] https://dre.pt/web/guest/legislacao-consolidada/-/lc/view?cid=74901117

incluindo som e imagem, relativa a uma pessoa singular identificada ou identificável («titular dos dados»); é considerada identificável a pessoa que possa ser identificada direta ou indiretamente, designadamente por referência a um número de identificação ou a um ou mais elementos específicos da sua identidade física, fisiológica, psíquica, económica cultural ou social". Particularmente relevante é também o Artigo 6.º "Condições de legitimidade do tratamento de dados". Neste contexto, refere-se ainda o artigo 35º da Constituição Portuguesa[37] - Utilização da Informática – que proíbe explicitamente, no artigo 3 o uso da informática para "tratamento de dados referentes a convicções filosóficas ou políticas, filiação partidária ou sindical, fé religiosa, vida privada e origem étnica, salvo mediante consentimento expresso do titular, autorização prevista por lei com garantias de não discriminação ou para processamento de dados estatísticos não individualmente identificáveis".

É portanto fundamental verificar a natureza dos dados de investigação recolhidos ou gerados no âmbito do trabalho científico e, se necessário, aplicar as necessárias práticas de anonimização de dados, de modo a cumprir a legislação em vigor.

[37] http://www.parlamento.pt/Legislacao/Paginas/ConstituicaoRepublicaPortuguesa.aspx

5 - Políticas e Diretrizes de Dados de Investigação

Dependendo da hierarquia em questão, uma política de dados poderá ser definida a nível institucional, a nível nacional, ou a nível do financiador. Indicam-se em seguida alguns exemplos relevantes neste contexto.

Políticas Institucionais

A Universidade de Cambridge apresenta um sumário das instituições que implementaram tais políticas, a nível internacional,[38] podendo as mesmas ser consultadas e acedidas livremente. De forma semelhante, o DCC apresenta informação detalhada acerca de vários financiadores a nível do Reino Unido.[39]

Fundação para a Ciência e Tecnologia (FCT)

A 5 de maio de 2014, a FCT publicou a "Política sobre a Disponibilização de Dados e outros Resultados de Projetos de I&D Financiados Pela FCT",[40] que consiste essencialmente num conjunto de orientações gerais e recomendações aos beneficiários de financiamento, acerca de práticas a ter em relação à produção, armazenamento e partilha de dados de investigação.

Políticas da União Europeia – Programa Horizonte 2020

Programa Horizonte 2020

No âmbito do programa Horizonte 2020, a União Europeia lançou o Piloto de Dados de Investigação Abertos[41] que visa melhorar e maximizar o acesso e a reutilização dos dados de investigação gerados por projetos financiados pela CE. A CE estabeleceu os requisitos para dados abertos de investigação, requerendo na sua política[42] a obrigatoriedade da publicação dos dados obtidos em acesso aberto: a) Dados para validar os resultados apresentados em publicações científicas; b) Outros dados, conforme especificado no plano de gestão de dados. Apenas mediante justificação específica e válida para projetos pontuais é aceitável o não cumprimento da partilha de dados.

---

[38] http://www.data.cam.ac.uk/funders
[39] http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies
[40] https://www.fct.pt/documentos/PoliticaAcessoAberto_Dados.pdf
[41] https://www.openaire.eu/opendatapilot
[42] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

6 - Recursos de Apoio

Nesta secção é identificado e indicado material de apoio aos conteúdos apresentados anteriormente.

i. Planeamento

Dá-se aqui destaque aos seguintes recursos, relevantes para a fase de planeamento:
- a já referida a ferramenta DMP-online[43] (DCC) e a informação disponibilizada no contexto do Piloto de Dados Abertos OpenAIRE,[44] para a criação de um plano de gestão de dados.
- os variados recursos existentes na plataforma FOSTER (Facilitate Open Science Training for European Research) na área de planos de gestão de dados[45]

- estimativa de custos: o projeto 4C (Collaboration to Clarify the Cost of Curation)[46] é um consórcio de vários parceiros, cujo foco é fornecer uma ferramenta para uma estimativa realista dos custos de preservação e curadoria digital.

ii. Metadados

Existem variados esquemas de metadados, sendo que estes podem ser divididos em esquemas interdisciplinares (geral) e esquemas disciplinares (específicos). Neste contexto dá-se destaque ao esquema de metadados Datacite,[47] interdisciplinar e interoperável com outros esquemas generalistas.
Para a escolha de um repositório que utilize um esquema de metadados específico de uma disciplina ou área de investigação a uma determinada disciplina sugerimos a consulta da base de dados re3data, e da procura pela disciplina desejada.

iii. Certificação de Repositórios

Como já foi referido, existem varias agências que fornecem certificados a repositórios digitais. Destacamos aqui as seguintes, por atuarem no contexto deste Kit:

- Data Seal of Approval (DSA)

[43] https://dmponline.dcc.ac.uk/
[44] https://www.openaire.eu/opendatapilot-dmp
[45] https://www.fosteropenscience.eu/taxonomy/term/139
[46] http://4cproject.eu/
[47] http://schema.datacite.org/

- Network of Expertise in long-term Storage and Accessibility of Digital Resources in Germany (NESTOR)
- German Institute for Standardization (DIN) standard 31644, Trustworthy Repositories Audit and Certification (TRAC) criteria
- International Organization for Standardization (ISO) standard 16363
- International Council for Science World Data System (ICSU-WDS) certification of WDS Members

iv. Repositórios de Dados

- Zenodo[48]

Zenodo é um repositório generalista, associado ao projeto OpenAIRE, pioneiro no acesso aberto e no movimento de dados abertos na Europa. Este projeto foi liderado pela comissão europeia que suportou o início de uma política de dados a nível Europeu. O CERN, enquanto parceiro do OpenAIRE suporta o repositório Zenodo em termos de infraestrutura, assegurando a capacidade de armazenamento e a preservação digital dos dados depositados.

- Dryad[49]

Dryad é um repositório de dados de investigação científica, sempre sujeitos a um processo de curadoria. Este repositório encontra-se integrado com variadas revistas científicas, cumprindo os requisitos das mesmas no que respeita à submissão conjunta de dados de investigação e facilitando o processo de revisão por pares. O depósito de dados tem um custo associado, dependendo do modelo financeiro e da revista associada à publicação.

- figshare[50]

Figshare é um repositório generalista parte do grupo empresarial Digital Science, uma empresa do grupo Macmillan Publishers. Figshare é um repositório online onde os institutos de investigação e investigadores podem preservar e publicar todo o tipo de conteúdos resultantes do processo de investigação. O seu uso é grátis e o depósito de dados inteiramente da responsabilidade do utilizador. Este repositório está integrado com variadas revistas científicas, o que não só aumenta significativamente a qualidade dos dados, como

[48] https://zenodo.org/
[49] http://datadryad.org/
[50] https://figshare.com/

facilita o processo de revisão por pares, enquanto garante a preservação dos respetivos dados.

- re3data[51]

O projeto re3data.org é um registo global de repositórios disciplinares focados em dados de investigação, sendo por isso, uma referência não só durante o decorrer do processo de investigação como durante a fase de planeamento de projeto. Permite encontrar repositórios por disciplina, país, tipo de conteúdo, assim como verificar variadas características de cada repositório encontrado. Este serviço serve como referência para institutos de investigação ou investigadores na escolha de um repositório adequado para os dados produzidos. Começou por ser um projeto financiado pela Fundação para a investigação na Alemanha tendo-se posteriormente fundido com o projeto DataCite, em 2015.

v. Conteúdos de formação generalistas

- Mantra

O projeto Mantra[52] (Universidade de Edimburgo) dispõe conteúdos online de forma gratuita, abordando o tema de dados e curadoria digitais e disponibilizando recursos específicos para estudantes, investigadores, académicos ou profissionais de informação.

- FOSTER

A plataforma Foster alberga vários conteúdos na área de ciência aberta, disponibilizando um vasto leque de recursos na área dos dados abertos[53]

- LIBER

O webinar[54] e o respetivo documento documento “23 coisas: Bibliotecas e Dados Científicos”[55] apresentam uma visão geral de recursos úteis e ferramentas online livres que podem servir para integrar a gestão de dados científicos no trabalho prático dos profissionais das bibliotecas.

[51] http://service.re3data.org/schema
[52] http://datalib.edina.ac.uk/mantra/
[53] https://www.fosteropenscience.eu/taxonomy/term/6
[54] https://www.youtube.com/watch?v=HGH6fVHrnKQ; http://libereurope.eu/
[55] https://b2share.eudat.eu/api/files/792ab2b7-8467-4a1a-af4b-f9a7491b07ba/23Things_Libraries_For_Research_Data_pt.pdf

vi. Serviços para Identificadores Persistentes

Existem vários serviços desenvolvidos sobre diferentes identificadores persistentes; Neste contexto destacamos os seguintes serviços:

- Resolução de Identificadores

Mediante a indicação de um número do identificador os serviços de resolução de Handles[56] e de DOIs[57] redirecionam para os metadados e conjunto de dados respetivo.

- Métricas de dados

Existem vários serviços que incluem de métricas de dados de investigação. Salientamos os serviços Data Citation Index[58] (Clarivate Analytics, antes pertencente à Thomson Reuters) e o PlumX,[59] que permitem estimar o impacto que conjuntos de dados publicados.

Uma boa referência acerca de métricas de dados de investigação é o artigo publicado pelo DCC.[60]

vii. Políticas de Dados Abertos

O projeto learn-RDM[61] fornece apoio na formulação e implementação de uma política de dados abertos, a nível institucional. Este projeto teve início em Fevereiro de 2016 estando prevista a publicação de um documento padrão para a implementação de uma política de dados de investigação a nível institucional.

O DISC (Data Information Specialists Committee - United Kingdom)[62] disponibiliza um guia bastante completo em relação à implementação de políticas, a nível institucional. Aborda variados temas, salientando considerações chave fundamentais à introdução de políticas sustentáveis de gestão de dados de investigação.

A Universidade de Cambridge apresenta um sumário das instituições que implementaram tais políticas, a nível internacional,[63] podendo as mesmas ser consultadas e acedidas livremente. De forma semelhante, o DCC apresenta informação detalhada acerca de vários financiadores a nível do Reino Unido.[64]

---

[56] https://hdl.handle.net/
[57] https://dx.doi.org/
[58] http://wokinfo.com/products_tools/multidisciplinary/dci/
[59] http://plu.mx/
[60] http://www.dcc.ac.uk/resources/how-guides/track-data-impact-metrics
[61] http://learn-rdm.eu/en/about/
[62] http://www.disc-uk.org/docs/guide.pdf
[63] http://www.data.cam.ac.uk/funders
[64] http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

viii. Normas de citação de Dados de Investigação

Com a crescente publicação independente de conjuntos dados torna-se fundamental, para a sua reutilização correta, a adoção de diretrizes e normas para a citação de dados. Esta é uma temática em crescimento e dependente do contexto em que os dados são usados.
O artigo publicado[65] no contexto do projeto CODATA será sem dúvida um bom ponto de partida, neste aspeto e questões relacionadas.

7 - Aplicação nos recursos RCAAP

O presente documento servirá de base ao curso de e-learning a disponibilizar no *site* para o efeito designado do RCAAP[66]. Este documento – enquanto documento PDF - reflete todos os conteúdos disponibilizados sendo no entanto complementado por atualizações colocadas *online*, no site de e-learning.

Divulgação

O Kit deverá ser divulgado juntamente da comunidade, nomeadamente fazendo não só uso dos contactos de e-mail dos participantes da Conferência de Dados de Investigação e do 1º e 2ºFórum-GDI, como também fazendo uso extensivo das redes sociais e blog-RCAAP.

Desenvolvimentos futuros

O Kit de Dados de investigação deverá ser revisto dentro do tempo máximo de 6 meses, após a sua publicação de forma a garantir a sua atualidade e relevância, no panorama nacional. Dentro desse espaço de tempo serão também adicionados alguns complementos ao Kit, nomeadamente:
- um questionário de modo a aferir a apreensão dos conteúdos apresentados
- um glossário de dados de investigação
- uma apresentação genérica (formato Power-Point) de modo a servir de material de formação para a comunidade

[65] http://datascience.codata.org/articles/abstract/10.2481/dsj.OSOM13-043/
[66] https://elearning.rcaap.pt/