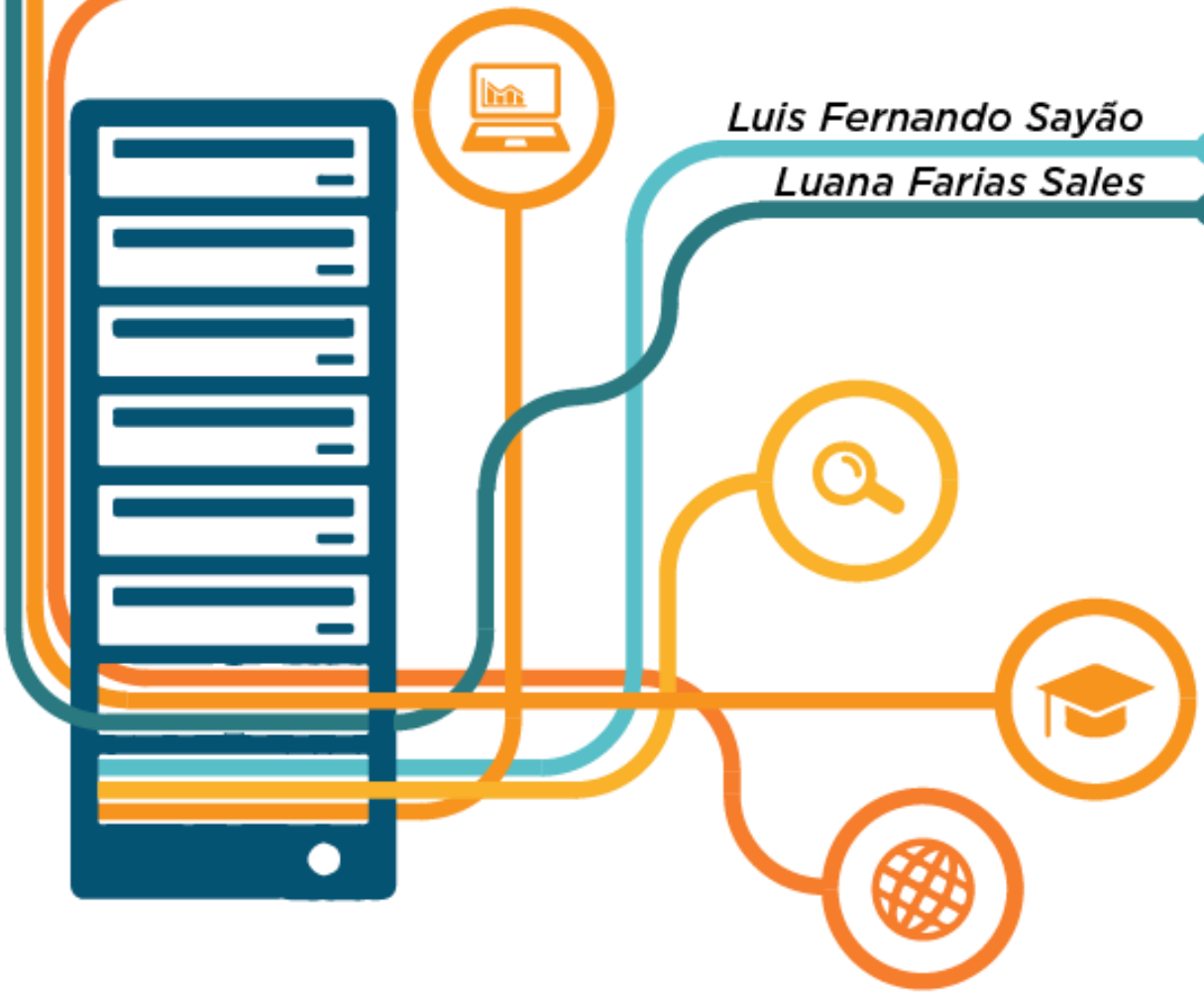
Guia de

Gestão de Dados de Pesquisa

para bibliotecários e pesquisadores

Luis Fernando Sayão

Luana Farias Sales



45 anos cin · ien · CNEN Comissão Nacional de Energia Nuclear · Ministério da Ciência, Tecnologia e Inovação · Governo Federal Brasil Pátria Educadora

COMISSÃO NACIONAL DE ENERGIA NUCLEAR

GUIA DE GESTÃO DE DADOS DE PESQUISA PARA BIBLIOTECÁRIOS E PESQUISADORES

LUIS FERNADO SAYÃO
Centro de Informações Nucleares

LUANA FARIAS SALES
Instituto de Engenharia Nuclear

CNEN

RIO DE JANEIRO

2015

Dados Internacionais de Catalogação na Publicação (CIP)

S274g
Sayão, Luis Fernando.
Guia de Gestão de Dados de Pesquisa para Bibliotecários e Pesquisadores / Luis Fernando Sayão, Luana Farias Sales. – Rio de Janeiro : CNEN/IEN, 2015.
90 p.

ISBN 978-85-61905-03-3

1. Dados de Pesquisa. 2. Gestão de dados de pesquisa. 3. Curadoria digital. I. Sales, Luana Farias. II. Título.
CDU

Sumário

1
INTRODUÇÃO

O reconhecimento do potencial informacional dos dados de pesquisa para a ciência contemporânea transforma a visão que os caracterizava como simples subprodutos dos processos de pesquisa. Naquele contexto, os dados eram considerados somente na sua configuração final, sem considerar os seus ciclos de vida, versões e linhagens e, via de regra, eram descartados ou armazenados em mídias ou em servidores sem a devida gestão quando os projetos eram concluídos. Quase sempre eram tragados silenciosamente pelo tempo: pela obsolescência tecnológica e pela fragilidade das mídias digitais[1].

Os pesquisadores, as instituições acadêmicas e as agências de fomento à pesquisa começam a compreender que esses dados, se devidamente tratados, preservados e gerenciados, podem constituir uma fonte inestimável de recursos informacionais para a pesquisa científica e para o ensino da ciência. Os repositórios de dados se incorporam rapidamente à infraestrutura mundial de informação científica e, dessa forma, as coleções de dados podem ser usadas, reusadas e compartilhadas. Potencialmente, esses dados podem capacitar os pesquisadores a formular novos tipos de indagações, hipóteses e a usar métodos analíticos inovadores no estudo de questões críticas para a ciência e para a sociedade[2].

Nessa direção, uma gestão eficiente dos dados é fundamental para o desenvolvimento de pesquisas de alta qualidade e excelência. A gestão de dados cobre todos os aspectos relativos à manipulação, organização, documentação e agregação de valor, e tem um papel crucial como facilitador nos processos de compartilhamento dos dados, na garantia da sustentabilidade e acessibilidade dos dados em longo prazo. As ações e comprometimentos promovidos pela gestão, coletivamente, permitem que os dados de valor possam ser reusados em outros projetos ao longo do tempo e do espaço[3].

A QUEM SE DIRIGE ESTE GUIA

A gestão de dados de pesquisa – pela amplitude do seu alcance na ciência contemporânea e pelo seu valor como recurso informacional – não é responsabilidade somente dos pesquisadores que criaram ou coletaram os dados. Muitas pessoas estão envolvidas nos processos de pesquisa e têm papeis importantes na garantia da qualidade, integridade, proveniência e preservação dos dados. Porém, o papel crucial ainda é do **pesquisador**.

---

[1] SAYÃO, Luís Fernando; SALES, Luana Farias. Dados abertos de pesquisa: ampliando os conceitos de acesso livre. **RECIIS – Rev. Eletron. de Comun. Inf. Inov. Saúde**. v. 8, n. 2, p. 76-92, 2014.

[2] BORGMAN, Cristine. Research data: who will share what, with whom, when, and why? In: CHINA--NORTH AMERICAN LIBRARY CONFERENCE, 5., 2010, Beijing. Disponível em: <http://works.bepress.com/borgman/238/>. Acesso em: 10 out. 2015.

[3] BALL, Ales. **A review of data management lifecycle models**. Batth, UK : University of Bath, 2012. Disponível em: <http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf>. Acesso em: 01 out. 2015.

Por outro lado, a **biblioteca de pesquisa** é a custodiante natural dos registros acadêmicos, e este papel se estende agora para incluir os dados de pesquisa. Os bibliotecários estão bem posicionados para trabalhar com os dados pelos seus conhecimentos em gestão de informação, metadados, descoberta de recursos, preservação digital, além disso, eles sempre estabeleceram uma relação longa e produtiva com os pesquisadores. Assim, este Guia se destina de forma privilegiada aos **pesquisadores** e **bibliotecários**.

Apesar do Guia ter sido elaborado no âmbito da Comissão Nacional de Energia Nuclear, o seu escopo de aplicação é geral, e pode ser utilizado em qualquer área, incluindo as áreas de ciências sociais.

OBJETIVOS DO GUIA

O objetivo do Guia é apresentar aos pesquisadores e bibliotecários os elementos básicos, conceitos, ferramentas, referências e melhores práticas para o planejamento da gestão de dados de pesquisa e para a efetiva ação ao longo de todo o ciclo de vida dos dados.

COMO CONSULTAR

O Guia pode ser lido sequencialmente e pode ser consultado como uma obra de referência para os interessados em tópicos específicos. Para tal, ele foi organizado de forma que a partir da compreensão global do Plano de Gestão de Dados (PGD) – CAPÍTULO 4 - os elementos que o compõe sejam capítulos que podem ser lidos de forma independente. Como ferramentas auxiliares, o Guia apresenta no **Apêndice I** um glossário relacionando aos principais conceitos necessários à compreensão da gestão de dados de pesquisa; no **Apêndice II** esquemas de metadados para dados de pesquisa; e no **Apêndice III** um índice remissivo das perguntas chave incluídas no Guia, que encurtam o caminho da consulta e orientam como proceder a cada passo ao longo do processo de gestão.

PRINCIPAIS FONTES

Este guia é amplamente baseado nas seguintes fontes:

- GREEN, Ann; MACDONALD Stuart; RICE, Robin. **Policy-making for research data in Repositories**: a guide. May 2009.
- ICPSR. **Guide to social science data preparation and archiving**. Ann Arbor: ICPSR, 2012.
- EYNDEN, Veerle et al. **Managing and data sharing**: best practice for researchrs. Colchester: UK Data Archive, 2011.

2
O QUE É DADO DE PESQUISA?

O Relatório da **Organização para a Cooperação e Desenvolvimento Econômico** (OCDE)[4], publicado em 2007, cujo título em português é "Princípios e diretrizes para o acesso a dados de pesquisa financiados por fundos públicos", descreve dados de pesquisa como "registros factuais usados como fonte primária para a pesquisa científica e que são comumente aceitos pelos pesquisadores como necessários para **validar os resultados do trabalho científico**" (p.13). O que se observa é que a amplitude do que se entende por dados de pesquisa sugere um conceito complexo que pode se manifestar numa multiplicidade de formas.

A noção de dados pode variar consideravelmente entre pesquisadores e, ainda mais, entre áreas do conhecimento. A constatação de que **os dados são gerados para diferentes propósitos, por diferentes comunidades acadêmicas e científicas e por meio de diferentes processos** intensifica ainda mais essa percepção de diversidade. Tipos de dados podem incluir, por exemplo, números, imagens, textos, vídeos, áudio, *software*, algoritmos, equações, animações, modelos, simulações. Alguns tipos de dados têm valor imediato e duradouro, enquanto outros adquirem valor ao longo do tempo; alguns dados são capturados num momento específico e irrecuperável, enquanto outros são passíveis de se reproduzir.

Essa **heterogeneidade intrínseca aos dados** de pesquisa implica que é necessário formular estratégias de **gestão de amplo espectro** que englobem os vários tipos de dados. O reconhecimento dessas diferenças torna-se crucial quando se estabelecem as opções gerenciais e tecnológicas para o **arquivamento persistente** e para a **curadoria digital** das coleções de dados de pesquisa.

QUAIS SÃO OS TIPOS DE DADOS DE PESQUISA?

Os dados de pesquisa podem ser caracterizados de várias formas, por exemplo, de acordo com sua **natureza**, **origem** ou de acordo com seu **status no fluxo de trabalho da pesquisa**. Cada uma dessas visões revela uma faceta desses recursos informacionais.

CLASSIFICAÇÃO DOS DADOS SEGUNDO A SUA ORIGEM[5]:

- **DADOS OBSERVACIONAIS**
  São dados obtidos **por meio de observações diretas**, que podem ser associadas a lugares e tempo específicos, como por exemplo, a erupção de determinado vulcão

[4] OECD. **OECD Principles and Guidelines for Access to Research Data from Public Funding.** OECD, 2007. Disponível em: <http://www.oecd.org/sti/sci-tech/38500813.pdf>. Acesso em: 01 out. 2015.
[5] GREEN, Ann; MACDONALD, Stuart; RICE, Robin. **Policy-making for research data in Repositories**: a guide. May 2009. Disponível em: <https://www.coar-repositories.org/files/guide.pdf>. Acesso em: 01 out. 2015.

numa data específica, a fotografia de uma supernova, o levantamento das atitudes de uma comunidade; os dados observacionais – por sua natureza instantânea – guardam uma importância crítica que os qualifica como registros históricos, pois **não podem ser coletados uma segunda vez** e, portanto, devem ser submetidos a processos de curadoria que os **preserve para sempre**.

- **DADOS COMPUTACIONAIS**
  São resultados da execução de modelos computacionais ou de simulações, seja, por exemplo, no domínio da física ou para a criação de ambientes virtuais culturais ou educacionais. Para esta categoria de dados a preservação por longo prazo pode não ser necessária, posto que os dados podem ser replicados ao longo do tempo. Entretanto, replicar o modelo ou a simulação no futuro pode exigir um grande número de informações que incluem descrição das dependências de *hardware*, *software* e outras dependências técnicas, e ainda os dados de entrada. É preciso notar que algumas vezes é mais conveniente preservar somente os dados de saída.

- **DADOS EXPERIMENTAIS**
  São **provenientes de situações controladas em bancadas de laboratórios**, como por exemplo, medidas de uma reação química. Em tese, dados experimentais provenientes "de experimentos que podem ser precisamente reproduzidos não necessitam ser armazenados indefinidamente; porém, na prática, nem sempre é possível reproduzir precisamente todas as condições experimentais, particularmente onde algumas variáveis experimentais não podem ser conhecidas e quando os custos de reprodução do experimento são proibitivos"[6].

A distinção definida por essa categorização é de grande importância na escolha das estratégias de arquivamento e preservação.

CLASSIFICAÇÃO DOS DADOS SEGUNDO A SUA NATUREZA:

- NÚMEROS, IMAGENS, VÍDEOS ou ÁUDIO, SOFTWARE, ALGORÍTIMOS, EQUAÇÕES, ANIMAÇÕES ou MODELOS e SIMULAÇÕES.

CLASSIFICAÇÃO DOS DADOS SEGUNDO A FASE DA PESQUISA[7]

- **DADOS BRUTOS, CRUS ou PRELIMINARES** (*RAW DATA* em inglês)
  São dados que vêm diretamente dos instrumentos científicos.

- **DADOS DERIVADOS**
  São resultados do processamento ou combinação de dados brutos ou de outros dados.

- **DADOS CANÔNICOS ou DADOS REFERENCIAIS**

[6] UK DATA ARCHIVE. **Create & manage data**: formatting your data. Disponível em: http://www.data-archive.ac.uk/create-manage/format>. Acesso em: 01 out. 2015.
[7] GREEN, Ann; MACDONALD, Stuart; RICE, Robin. **Policy-making for research data in Repositories**: a guide. May 2009. Disponível em: <https://www.coar-repositories.org/files/guide.pdf>. Acesso em: 01 out. 2015.

São coleções de dados consolidados e arquivados geralmente em grandes centros de dados, por exemplo, sequência genética, estrutura química, etc.

Muitas áreas de pesquisa fazem uso também de **dados produzidos por órgãos do governo**. Embora estes dados não tenham sido originalmente coletados para fins de pesquisa, eles se tornam dados de pesquisa uma vez que tenham sido modificados, processados ou expandidos.

3
CICLO DE VIDA DOS DADOS DE PESQUISA

Os dados e as coleções de dados de pesquisa possuem um tempo de vida maior que os projetos de pesquisa que os criaram. Isso significa que pesquisadores, professores, estudantes e outros profissionais podem continuar a trabalhar sobre esses dados após os projetos e financiamentos tenham sido cessados. Novos projetos de pesquisa podem analisar ou adicionar novos elementos a esses dados de forma que eles possam ser reusados por outros pesquisadores, reiniciando um novo ciclo.

Há uma série de concepções de modelos de ciclo de vida de dados de pesquisa, cada um com particularidades e objetivos determinados, muitas vezes orientados para domínios de conhecimentos específicos. A importância desses modelos é que eles oferecem uma estrutura que representa as muitas operações que precisarão ser realizadas sobre os registros de dados durante a sua vida, garantido que eles possam ter a sua usabilidade otimizada e estendida.

Há alguns modelos que se tornaram referências para pesquisadores, bibliotecários e gestores de dados, são eles[8]:

- **DIGITAL CURATION CENTRE (DCC) CURATION LIFECYCLE MODEL**[9]
- **DATAONE DATA LIFECYCLE**[10]
- **DDI COMBINED LIFECYCLE MODEL**[11]
- **UK DATA ARCHIVE DATA LIFECYCLE**[12]

Para o propósito do presente Guia, tomaremos como referência o ciclo de vida definido pelo **DataONE**, por estar mais próximo dos objetivos do documento. Este ciclo de vida tem oito etapas:

PLANEJAR
COLETAR
ASSEGURAR A QUALIDADE
DESCREVER
PRESERVAR
DESCOBRIR
INTEGRAR
ANALISAR

[8] BALL, Ales. **A review of data management lifecycle models**. Batth, UK: University of Bath, 2012. Disponível em: <http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf>. Acesso em: 01 out. 2015
[9] <http://www.dcc.ac.uk/resources/curation-lifecycle-model>
[10] <http://escholarship.org/uc/item/7tf5q7n3#page-1>
[11] <http://www.ddialliance.org/Specification/DDI-Lifecycle/>
[12] <http://www.data-archive.ac.uk/create-manage/life-cycle>

‘

QUAIS SÃO AS ETAPAS DO CICLO DE VIDA DOS DADOS DE PESQUISA?[13]

- **PLANEJAR**
  Descrição dos dados que serão compilados e como eles serão gerenciados e disponibilizados para acesso durante o seu tempo de vida.

- **COLETAR**
  As observações são realizadas manualmente ou por sensores ou outros instrumentos, e os dados são colocados em formas digitais.

- **ASSEGURAR**
  A qualidade dos dados é assegurada por meio de controles e inspeção.

- **DESCREVER**
  Dados são descritos de forma acurada usando padrões de metadados apropriados.

- **PRESERVAR**
  Dados são submetidos a um arquivo apropriado (centro de dados) para preservação de longo prazo.

- **DESCOBRIR**
  Dados potencialmente úteis são descobertos e acessados juntamente com informações relevantes sobre os dados (metadados).

- **INTEGRAR**
  Dados de diversas fontes são combinados para formar um conjunto de dados homogêneo que pode ser prontamente analisado.

- **ANALISAR**
  Dados são analisados.

[13] STRASSER, Carly et al. **Primer on Data Management**: What you always wanted to know. California: CDL, 2012. Disponível em: <http://escholarship.org/uc/item/7tf5q7n3#page-1>. Acesso em: 01 out. 2015.

TODAS AS ETAPAS DO CICLO DE VIDA TÊM QUE SER CUMPRIDAS?

Um pesquisador ou uma equipe de pesquisadores está normalmente engajada em todos os aspectos do ciclo de vida dos dados, no papel de criador e também como usuário dos dados. Algumas equipes de cientistas – por exemplo, aquelas vinculados à modelagem e sínteses – podem criar novos dados no processo de descobrir, integrar, analisar e sintetizar dados existentes.

Entretanto, alguns projetos podem usar apenas parte do ciclo de vida, por exemplo, um projeto envolvido com **meta-análise** pode se concentrar nas etapas **descobrir, integrar** e **analisar** e desconsiderar as outras etapas, ou seja, alguns projetos podem não seguir de forma linear o caminho delineado pelo modelo.

4

PGD - PLANO DE GESTÃO DE DADOS

Para a efetiva gestão de dados de pesquisa, o planejamento é uma fase essencial. Ele se inicia quando a pesquisa ainda está sendo delineada e deve considerar como os dados serão gerenciados durante o desenvolvimento do projeto e como eles serão compartilhados depois. Dessa forma é necessário formalizar as ações e compromissos que serão estabelecidos em relação aos dados desde os seus primeiros estágios.

O PGD descreve o ciclo de vida de gestão para todos os dados que serão coletados, processados ou gerados por um projeto de pesquisa. De uma forma abreviada, ele se constitui em um documento formal que estabelece um compromisso de como esses dados serão tratados durante todo o desenvolvimento do projeto, e também após a sua conclusão.

Para isso, o PGD descreve, de uma forma geral, que dados serão processados, coletados ou gerados; quais as metodologias e padrões que serão utilizados nesses processos; se, como e sob que condições esses dados serão compartilhados e/ou tornados abertos para a comunidade de pesquisa; e como eles serão curados e preservados.

Posto que o PGD espelha uma situação dinâmica, é necessário observar que ele **não é um documento fixo no tempo**, ao contrário, ele se desenvolve e ganha mais precisão e solidez durante o tempo de vida do projeto[14]

No contexto atual, caracterizado pela riqueza de dados, o PGD se torna rapidamente um documento essencial no cotidiano dos pesquisadores, posto que, nos últimos anos, muitas **agências financiadoras de pesquisa** têm introduzido no seu elenco de exigências para financiamento de projetos de pesquisa que um **plano de gestão e de compartilhamento de dados faça parte dos pedidos de auxílio**.

Entretanto, o PGD não é um documento burocrático e sua elaboração não deve ser pensada como uma mera tarefa administrativa na qual um texto padronizado possa ser utilizado para todos os projetos. **Ele deve ser tratado como uma carta de intenções que considere o que realmente é necessário para a preservação, compartilhamento e reuso dos dados**.

[14] EUROPEAN COMISSION. **Guidelines on data management in horizon 2020**. Dec. 2013. Disponível em: <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf>. Acesso em: 01 out. 2015.

A pressão por cumprir cronogramas apertados e o custo da gestão de dados são fatores críticos no cumprimento do PDG, portanto você deve planejar a gestão de dados de acordo com as necessidades e propósito de sua pesquisa. Muitos aspectos dessa gestão podem ser incorporados nos fluxos normais da coordenação, gestão e procedimentos do projeto de pesquisa, ou seja**, as medidas de gestão dos dados devem fazer parte do fluxo de trabalho da sua pesquisa**.

POR QUE CRIAR UM PLANO DE GESTÃO DE DADOS DE PESQUISA?

Há muitas razões para que seja necessária a elaboração de um plano de gestão de dados, porém a mais importante é que a gestão de dados é uma das áreas essenciais na **conduta responsável da pesquisa nos ambientes científicos atuais**. Além do mais, auxilia os pesquisadores a considerar, **ainda na fase de concepção e planejamento do projeto de pesquisa**, como os dados serão **geridos durante a pesquisa** e como serão **posteriormente preservados e compartilhados** com a comunidade científica mais ampla[15].
As principais razões para a criação de um PGD são as seguintes:

- Ajustar o seu projeto de pesquisa às **políticas mandatórias** da sua instituição e/ou dos órgãos de fomento à pesquisa;
- Assegurar a **integridade** da pesquisa e o seu **potencial de replicação**;
- Assegurar que os dados e demais registros de pesquisa sejam **acurados, completos, autênticos e confiáveis**;
- Aumentar a sua **eficiência como pesquisador** – um plano que organize os dados e seu armazenamento permite que você foque na sua pesquisa. Você estará mais capacitado a localizar e usar os seus dados e compartilhá-los com os seus colaboradores;
- Permitir que os seus dados sejam **compreensíveis agora e no futuro** – se os dados são bem documentados antes e durante a formação da coleção de dados, eles serão mais facilmente **entendidos** e **reutilizados;**
- **Economizar tempo e recursos** a longo prazo;
- Aumentar a **segurança dos dados** e minimizar os **riscos de perda;**
- Evitar a **duplicação de esforços** na recoleta ou regeração dos dados, possibilitando que outros pesquisadores se beneficiem dos seus dados e os interprete em outros contextos e com novas visões;

[15] EYNDEN, Veerle et al. **Managing and data sharing:** best practice for researchrs. Colchester: UK Data Archive, 2011. Disponível em: <http://www.data-archive.ac.uk/media/2894/managingsharing.pdf>. Acesso em: 01 out. 2015.

- **Aumentar a visibilidade da pesquisa** – se os seus dados foram planejados para estarem organizados e corretamente arquivados, eles poderão ser identificados, recuperados e citados, aumentando a visibilidade da sua pesquisa e o seu prestígio como pesquisador;

- **Tornar mais fácil a preservação e o arquivamento** – ajustando antecipadamente a geração de dados tomando como referência as práticas, formatos e demais padrões mais adequados ao arquivamento e à preservação de longo prazo, torna a gestão de dados mais fácil e menos custosa; além do mais, tornam os dados mais aderentes aos requisitos dos repositórios e centros de dados.

4.1
COMO CRIAR UM PLANO DE GESTÃO DE DADOS?

O PGD pode ser criado por meio de padrões e *templates* definidos pela sua instituição ou pelas agências de financiamento que patrocina o seu projeto, há ainda ferramentas *on-line* que podem auxiliar você na elaboração do seu plano. Abaixo é apresentado um formato geral que requisita as informações mais comuns presentes nos PGDs. Caso a sua instituição não tenha ainda um modelo próprio, você pode seguir o roteiro abaixo. Ele é fortemente baseado nos elementos recomendados pelo DataONE e por outras organizações importantes como o JISC[16], o DCC[17] e o ICPSR[18].

4.1.1
INFORMAÇÕES SOBRE OS DADOS: TIPOS, VOLUME, PROCESSAMENTO, FORMATOS, ARQUIVAMENTO...

A pesquisa científica produz e coleta dados que são muito variados e heterogêneos e que têm natureza, formatos diferentes e são coletados em volumes variados e passam por diferentes processos que dependem de cada disciplina e dos objetivos da pesquisa, portanto é necessário descrever, com algum grau de detalhe, as principais características desses dados, incluindo a natureza e origem, escopo e a escala dos dados que serão produzidos. Isto vai ajudar os revisores e outros pesquisadores a compreenderem os dados, sua relação com os dados existentes e os possíveis riscos de disseminá-los[19].

[16] <https://www.jisc.ac.uk/>
[17] <http://www.dcc.ac.uk/>
[18] <https://www.icpsr.umich.edu/icpsrweb/landing.jsp>
[19] ICPSR. **Guide to Social Science Data Preparation and Archiving**. 2012. Disponível em: <http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf>. Acesso em: 10 out. 2015.

QUE TIPO DE DADOS SUA PESQUISA VAI PRODUZIR?

Liste os dados que seu projeto irá produzir e **os caracterize** em termos de natureza, origem e processamento: eles podem ser observacionais, experimentais, brutos ou derivados, simulações, coleções físicas, modelos, *software*, imagens, vídeos e muito mais.

QUE QUANTIDADE DE DADOS SERÁ GERADA PELA PESQUISA?

Com base na sua hipótese e no plano de amostragem avalie o **volume de dados** que o seu projeto irá gerar

COMO OS DADOS SERÃO COLETADOS?

Você deve especificar também os **métodos como os dados serão adquiridos**, isto inclui informações sobre quem, o que, quando e onde (como as amostras serão coletadas e analisadas? Que instrumentos serão usados?)

COMO OS DADOS SERÃO PROCESSADOS?

Uma vez que os dados foram adquiridos, deve ser especificado como eles serão processados. "Esta etapa deve ser considerada antes do projeto, pois pode afetar a maneira como os dados serão organizados, quais formatos serão usados, e quanto deve ser previsto, em termos orçamentários, para hardware e software. Devem ser considerados neste momento itens como que *software* poderá ser usado, que algoritmos serão empregados, e como esses itens se enquadram no fluxo de trabalho do projeto"[20]

QUAIS OS FORMATOS DE ARQUIVO QUE SERÃO USADOS?

Os formatos de arquivo dos dados que você planeja usar devem ser declarados e sua escolha deve ser justificada. Descreva os formatos nas fases de submissão, distribuição e preservação, observando que esses formatos podem ser os mesmos. Na sua escolha você deve considerar os padrões que são usados na sua área de pesquisa. Se os dados forem arquivados por longo prazo, é necessário considerar o uso de formatos padronizados e não proprietários, que são mais fáceis de serem interpretados no futuro, de forma independente de plataforma tecnológica (*hardware* e *software*).
(Considere se um banco de dados relacional ou outra estratégia de organização de dados pode ser mais apropriado para a sua pesquisa)

[20] STRASSER, Carly et al. **Primer on Data Management**: What you always wanted to know. California: CDL, 2012. Disponível em: <http://escholarship.org/uc/item/7tf5q7n3#page-1>. Acesso em: 01 out. 2015.

COMO OS ARQUIVOS SERÃO NOMEADOS?

É importante descrever também a convenção **adotada para dar nomes para seus conjuntos de dados, arquivos e pastas**. Convencionando isso de antemão, você estará menos propenso a mudar ou reorganizar os arquivos durante o projeto.

QUAIS SÃO AS MEDIDAS DE GARANTIA E CONTROLE DE QUALIDADE?

Você deve identificar quais são as medidas que você planeja adotar para **garantir e controlar a qualidade dos dados**; é necessário incluir também o que será feito durante e depois dos dados coletados, e ainda no curso da análise dos dados.

HÁ COLEÇÕES DE DADOS DISPONÍVEIS QUE SERVEM PARA SUA PESQUISA?

A **revisão dos dados existentes em periódicos e arquivos de dados** da sua área de pesquisa reforçará o valor de seu projeto e justificará mais claramente por que os dados atualmente disponíveis são inadequados para responder as suas questões de pesquisa.

SERÃO USADOS DADOS JÁ EXISTENTES?

Se **dados já existentes podem ser usados na sua pesquisa**, identificá-los e determinar suas origens (proveniência) é uma informação importante e deve ser registrada, bem como a relação entre esses dados e os dados que você está coletando. Se a sua coleção de dados será combinada com os dados já existentes, cabe definir como será **assegurada a compatibilidade de formatos**.

COMO OS DADOS SERÃO MANTIDOS A CURTO PRAZO?

Você precisa descrever **como os dados serão gerenciados logo após o termino do projeto**; isto significa planejar como manter o controle sobre as diferentes versões dos seus dados e das análises; como você fará *backup* de seus dados; se há computadores destinados a isso na sua instituição. Considere as opções de *backup* na sua instituição (*on-site*) e externamente (*off-site*). Descreva a sua estratégia para garantir a segurança dos dados, especialmente no caso de **dados sensíveis**. Delineie os possíveis usuários dos dados.

QUEM SERÁ O RESPONSÁVEL PELA GESTÃO DE CURTO PRAZO?

Identifique quem são os responsáveis pela gestão de curto prazo na sua instituição; determine papéis e responsabilidades para a gestão, arquivamento, controle de versões e procedimentos de *backup*.

4.1.2
METADADOS

Uma documentação exaustiva dos dados é a chave para a **compreensão do significado deles agora e no futuro.** Sem uma descrição minuciosa do contexto tecnológico dos arquivos de dados, do contexto no qual os dados foram criados ou coletados, das medidas que foram feitas, dos detalhes espaciais e temporais, dos instrumentos usados, dos parâmetros e unidades e da qualidade dos dados e da sua proveniência, é improvável que os dados possam ser descobertos, interpretados, gerenciados e efetivamente usados e reusados. **Os metadados cumprem essa tarefa, porque eles são a documentação dos dados**. Os metadados que são usados para descreverem os dados permitem que eles estejam autodocumentados agora e no futuro[21].

Nessa direção é importante que você delineie **os metadados que serão utilizados para descrever os dados que serão gerados/coletados por sua pesquisa**. Como os metadados são normalmente a única forma de comunicação entre os produtores de dados e as análises secundárias, metadados de qualidade são essenciais para o efetivo uso dos dados[22,23].

QUE METADADOS SÃO NECESSÁRIOS? Neste momento você tem que definir qual é o **elenco de metadados que são necessários** para que os **dados possam ter significado** e possam ser interpretados ao longo do tempo e do espaço.
COMO OS METADADOS SERÃO CRIADOS E/OU CAPTURADOS? Você deve informar também no seu Plano de Gestão de Dados **como os metadados serão criados ou capturados**. Por exemplo, seu caderno de campo ou de laboratório será usado para registrar as informações críticas? Instrumentos tais como unidades de GPS serão aperfeiçoadas para a coleta de dados? Os metadados serão salvos automaticamente pelos instrumentos que você está usando? Os dados precisarão de outros profissionais, como bibliotecários, para serem descritos?

[21] SURA. **A Step-By-Step Guide to Data Management.** August 2013. Disponível em: <http://www.lib.ua.edu/wiki/sura/index.php/A_Step-By-Step_Guide_to_Data_Management>. Acesso em: 01 out. 2015.
[22] ICPSR. **Guide to Social Science Data Preparation and Archiving**. 2012. Disponível em: <http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf>. Acesso em: 10 out. 2015.
[23] DATAONE. **Tutorials on data management Lesson 03**: Data Management Planning. Disponível em: <https://www.dataone.org/sites/all/documents/L03_DataManagementPlanning.pptx>. Acesso em: 10 out. 2015.

QUE ESQUEMA OU PADRÃO DE METADADO SERÁ USADO?

Informe que esquema (ou formato) de metadados você vai usar para descrever os seus dados. Existem muitos e diferentes padrões de metadados, alguns orientados para disciplinas especificas e outros de aplicação mais geral; consulte a sua biblioteca ou os seus colegas para saber qual o esquema que é mais adequado para a sua pesquisa. Caso não exista esquema que se adeque à sua disciplina, delineie os elementos que você acha importantes para descrever os seus dados.

Considere também verificar junto ao repositório ou centro de dados em que você pretende arquivar os seus dados as **exigências que eles têm sobre que esquema de metadados** aplicar.

Justifique no Plano a escolha do esquema de metadados, considerando a **sua comunidade de pesquisa**, o **repositório ou centro de dados** que você usará para arquivar os dados e a natureza do **seu projeto**.

4.1.3
POLÍTICA DE ACESSO, COMPARTILHAMENTO E REUSO

O acesso e compartilhamento de dados de pesquisa **contribui de forma significativa para que** a **ciência avance** e **maximize os investimentos aplicados em programas de pesquisa**. Estudos recentes concluem que o nível de produtividade da atividade de pesquisa aumenta e que o número de publicações baseadas em dados de pesquisa cresce dramaticamente quando os dados de pesquisa são compartilhados por meio de repositórios e arquivos de dados. Pesquisadores que compartilham seus dados publicamente recebem mais citações[24].

A importância das informações presentes nesta seção reside no fato de que a sua instituição e as agências que financiam sua pesquisa precisam saber se você está planejando preparar os seus dados para o compartilhamento com as comunidades potencialmente interessadas, em um tempo razoável, após a conclusão do seu projeto.

Contudo, é preciso **atentar para as restrições** que determinam se um conjunto de dados pode ser disponibilizado abertamente – parcial ou completamente - para compartilhamento com outros pesquisadores, posto que nem todos os dados gerados por pesquisa podem ser livremente distribuídos. Para dados de natureza sensível, por exemplo, que envolvem seres humanos, processos patenteáveis, espécies e ambientes ecológicos em risco, etc., em que o acesso público não é apropriado, você deve indicar que medidas especiais devem ser tomadas para acesso aos dados (por exemplo, acordo

[24] PEEERJ. **Scientists who share data publicly receive more citations**. October 2013. Disponível em: <http://www.eurekalert.org/pub_releases/2013-10/p-sws092413.php>. Acesso em: 10 out. 2015.

de consentimento informando anonimização dos dados, acesso unicamente por meio de redes seguras).

Duas questões críticas devem ser consideradas no momento do planejamento da gestão de dados:

- **PROTEÇÃO DOS DADOS**:
  Conjunto de dados onde pessoas, agregados familiares ou empresas são identificados. Nesse caso, via de regra, não é possível publicar os dados abertamente, mas em algumas situações, versões dos dados que passem por processos de anonimização podem ser passíveis de disseminação.

- **DATABASE *COPYRIGHT***:
  Quando os dados de pesquisa são derivados e/ou elaborados a partir de uma base de dados comercial pré-existente. Nesse caso, normalmente não é possível republicar as extrações significantes devido a questões de *copyright* e termos legais de uso.

É importante observar que mesmo quando os dados são disponibilizados abertamente, nem sempre é possível republicá-los livremente.

QUAIS SÃO AS OBRIGAÇÕES DE COMPARTILHAMENTO?[25] Relate as obrigações que você em relação ao compartilhamento dos seus dados; **políticas mandatórias para compartilhamento de dados podem vir da sua instituição, da agência financiadora, ou da sociedade científica** a que você está associado. **Existem também obrigações legais para compartilhamento dos dados**.
COMO OS DADOS SERÃO COMPARTILHADOS? Você também deve descrever os detalhes de **como você irá compartilhar os seus dados**: quanto tempo depois dos dados coletados eles estarão disponíveis para os seus colegas? Quando o aceso será aberto para todos os usuários interessados? Quem acessará esses dados? Como os dados serão acessados? Em qual repositório os dados serão armazenados? Que tipo de repositório? O coletor, o criador e o líder do projeto terão direitos exclusivos sobre os dados durante certo período de tempo (período de embargo)?

[25] DATAONE. **Tutorials on data management Lesson 03**: Data Management Planning. Disponível em: <https://www.dataone.org/sites/all/documents/L03_DataManagementPlanning.pptx>. Acesso em: 10 out. 2015.

HÁ QUESTÕES ÉTICAS E DE PRIVACIDADE ASSOCIADAS AOS DADOS?

Você deve se assegurar de que as **questões éticas e de privacidade dos seus dados serão corretamente endereçadas**. Se os seus dados envolvem, por exemplo, seres humanos, espécies em risco, ou habitats sensíveis, você deve tomar medidas especiais quando do compartilhamento dos dados.

HÁ QUESTÕES ASSOCIADAS À PROPRIEDADE INTELECTUAL E COPYRIGHT?

Descreva as **questões de propriedade intelectual e *copyright*** associados aos seus dados: A quem pertence o *copyright* de seus dados? Informe se os direitos serão transferidos para outra organização para distribuição e arquivamento; se algum material sujeito a *copyright* (por exemplo, instrumentos ou escalas) for usado, informe como o projeto irá obter permissão para usar e disseminar esse material.
Existem também outras considerações relativas a essa questão como **período de embargo** sobre dados que envolvem patentes, políticas e exigências de periódicos científicos.

QUAIS SÃO OS USOS FUTUROS E OS USUÁRIOS POTENCIAIS DOS MEUS DADOS?

Delineie os possíveis usos futuros dos seus dados e os usuários potenciais, essa reflexão ajuda a **determinar o repositório de dados mais apropriado para arquivar a sua coleção de dados**.

COMO OS DADOS PODEM SER CITADOS?

É importante descrever também como os seus **dados deverão ser citados** quando eles forem usados. Uma medida concreta é **atribuir um identificador persistente aos seus dados**, como por exemplo, o DOI (*Digital Object Identifier*)[26].

4.1.4
GESTÃO DO ARQUIVAMENTO DE LONGO PRAZO: PRESERVAÇÃO DIGITAL DOS DADOS DE PESQUISA

O compartilhamento e reuso dos dados de pesquisa, assim como a formação da memória digital das instituições de pesquisa, implica a necessidade de que os dados de pesquisa gerados e coletados sejam depositados em ambientes que garantam sua preservação ativa por longo prazo, mantendo as suas características de autenticidade,

[26]<https://www.doi.org/>

integridade e proveniência, de forma que eles estejam sempre disponíveis e prontos para serem usados.

Conteúdos digitais exigem ações de preservação constantes para que permaneçam viáveis – isto é, que possam ser lidos a partir de uma mídia digital – e interpretáveis.

Portanto, você deve informar no Plano de Gestão de Dados como você pretende fazer a **gestão de longo prazo dos dados**. Existem várias opções que podem ser utilizadas para esta fase dos dados, elas incluem **repositório institucional de sua instituição, repositórios associados aos periódicos científicos e repositórios e centros de dados que se dedicam a disciplinas específicas.**

Os repositórios e centros de dados são as opções mais adequadas, caso haja arquivos dessa natureza compatíveis com os dados da sua área de pesquisa, posto que eles podem assegurar que os dados serão curados e manipulados de acordo com as boas práticas da preservação digital[27]. Esses arquivos podem ainda oferecer orientações sobre como preparar os metadados, como preservar os dados, que formatos de arquivos usar e como disponibilizar serviços adicionais aos futuros usuários de seus dados. Os centros de dados podem – em continuidade a sua missão - oferecer ferramentas que apoiem a descoberta, o acesso e a disseminação de dados em resposta às necessidades dos usuários[28].

QUE DADOS SERÃO PRESERVADOS?

Nem todos os dados precisam ser preservados, por isso você deve, em primeiro lugar, selecionar os dados que passarão por processo de gestão de longo prazo. Em geral todos os **dados brutos** devem ser mantidos; todo produto que se configure como dado de pesquisa que tenha exigido tempo e muitos recursos para ser obtido deve ser preservado. **Qualquer dado que não pode ser facilmente substituído deve ser preservado**.

ONDE OS DADOS SERÃO ARQUIVADOS?

Depois você deve identificar onde os seus dados serão **arquivados para a gestão de longo prazo**. Uma medida importante é **identificar os repositórios ou centro de dados** mais comumente usados pela sua área de pesquisa, esses arquivos são mais duradouros e seguros do que *website* pessoal ou do seu laboratório. Verifique se sua instituição possui um repositório que aceita a submissão de dados de pesquisa.

[27] ICPSR. **Guide to Social Science Data Preparation and Archiving**. 2012. Disponível em: <http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf>. Acesso em: 10 out. 2015.
[28] STRASSER, Carly et al. **Primer on Data Management**: What you always wanted to know. California: CDL, 2012. Disponível em: <http://escholarship.org/uc/item/7tf5q7n3#page-1>. Acesso em: 01 out. 2015.

É NECESSÁRIO CONVERTER OS FORMATOS DOS DADOS?

O seu plano precisa descrever também que conversões de **formatos serão necessárias para garantir usabilidade dos dados no futuro**. Contate, logo nas fases iniciais do projeto, o repositório de dados ou centro de dados que você irá usar para arquivar os seus dados, assim você terá certeza de que eles serão criados no formato correto e recomendado para arquivamento de longo prazo, isso economizará um bom tempo em operações de conversão mais tarde.

QUEM SERÁ O RESPONSÁVEL PELO CONTATO COM O CENTRO DE DADOS?

Indique a pessoa que será responsável por manter contato com o centro de dados, isso será particularmente importante se existirem restrições de uso para os dados. Por exemplo, a exigência de que o usuário potencial faça contato com o coletor de dados antes de reusá-los.

4.1.5
ORÇAMENTO: CUSTOS ENVOLVIDOS NA GESTÃO DE DADOS

As atividades de gestão e compartilhamento de dados necessitam ser orçadas dentro do projeto de pesquisa em termos de tempo e de recursos. O pesquisador deve estimar os custos relativos à preparação dos dados, incluindo a documentação, para o compartilhamento e arquivamento. Algumas atividades potencialmente custosas – em termos de dinheiro e tempo - são listadas abaixo[29]:

- Preparação de documentação de alta qualidade;
- Ações relativas às questões de confidencialidade e do consentimento informado;
- Preparação e seleção de material para depósito.
-

QUE CUSTOS DEVEM SER PREVISTOS?

Considere no seu orçamento itens tais como **custos de homem-hora seu e de especialistas contratados na preparação dos dados e documentação**, requisitos de

[29] ICPSR. **Guide to Social Science Data Preparation and Archiving**. 2012. Disponível em: <http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf>. Acesso em: 10 out. 2015.

hardware e de pessoal necessários para o tratamento dos dados, bem como os custos associados com o arquivamento dos dados.

COMO ESSES CUSTOS SERÃO PAGOS?

Informe também como serão cobertos os custos associados à gestão dos dados. De maneira ideal, eles devem estar previstos nos pedidos de auxílio submetidos às agências de fomente.

RESUMO DAS INFORMAÇÕES RECOMENDADAS PARA COMPOR O PLANO DE GESTÃO DE DADOS	
DESCRIÇÃO DOS DADOS	• Tipo de dados produzidos pela pesquisa • Quantidade de dados que será coletada • Como os dados serão coletados • Como os dados serão processados • Formatos de arquivo que serão usados • Como os arquivos serão nomeados • Medidas para garantir a qualidade dos dados • Coleções de dados disponíveis • Dados existentes que serão usados • Preservação de curto prazo • Responsáveis pela gestão de curto prazo
METADADOS	• Metadados necessários • Como os metadados serão criados • Esquema que será usado
POLÍTICA DE ACESSO, COMPARTILHAMENTO E REUSO	• Obrigações de compartilhamento • Como os dados serão compartilhados • Questões éticas e de privacidade • Propriedade intelectual e copyright • Usos futuros e usuários potenciais • Citação dos dados
GESTÃO DO ARQUIVAMENTO DE LONGO PRAZO: PRESERVAÇÃO DIGITAL DOS DADOS DE PESQUISA	• Que dados serão preservados • Onde os dados serão arquivados • Necessidade de formatação dos dados • Responsável pelo contato com o centro de dados
ORÇAMENTO: CUSTOS ENVOLVIDOS NA GESTÃO DE DADOS	• Custos previstos • Como os custos serão cobertos

5
DOCUMENTE SEUS DADOS

PARA SER COMPARTILHADOS E USADOS, OS SEUS DADOS PRECISAM ESTAR BEM DESCRITOS.

METADADO É UMA FERRAMENTA IMPORTANTE PARA UMA DESCRIÇÃO PADRONIZADA DOS DADOS

Uma parte de grande importância na Gestão de Dados de Pesquisa é assegurar que os dados possam ser **compreendidos e interpretados por qualquer usuário agora e no futuro**. Isto exige uma **descrição clara e detalhada dos dados**, além de anotações adicionais e informações contextuais que possibilitem que os dados transmitam informação e conhecimento no tempo e no espaço. Isto é efetivado pela documentação que deve acompanhar os dados, ou seja, a **DOCUMENTAÇÃO DOS DADOS**.

A documentação que acompanha os dados **explica como estes recursos foram coletados ou gerados, o que os dados significam, qual é o seu conteúdo e estrutura, quais foram as manipulações a que eles foram submetidos.** Documentar os dados é considerado uma das melhores práticas na criação, organização e gestão de dados, além de ser uma estratégia importante para a preservação digital dos dados[30].

Dessa forma, para que os seus dados possam ser identificados, encontrados, acessados, usados e reusados de maneira apropriada por pesquisadores ou outros possíveis interessados, seus **dados devem estar acompanhados de uma documentação completa que descreva todos os seus aspectos**. Dados bem documentados têm mais chances de serem descobertos na Web, citados por terceiros e terem seu valor creditado aos autores.

Uma parte importante da documentação é formada por **metadados**. Usá-los torna mais **fácil achar e usar os dados ao longo do tempo**.

Metadado é um subconjunto padronizado e estruturado da documentação dos dados, formado por elementos de informação bem definidos – por exemplo, “titulo”, “autor”, “resumo”, “fonte” - que ajudam a **conferir contexto e informar a proveniência dos seus dados**, ou seja, a procedência e o histórico desses dados para pessoas e sistemas. Nessa direção, os metadados informam sobre a origem, propósito, tempo de referência,

[30] EYNDEN, Veerle et al. **Managing and data sharing**: best practice for researchrs. Colchester: UK Data Archive, 2011. Disponível em: <http://www.data-archive.ac.uk/media/2894/managingsharing.pdf>. Acesso em: 01 out. 2015.

localização geográfica, autor, condições de acesso e termos de uso das coleções de dados e muito mais.

Portanto, **produzir uma boa documentação significa principalmente atribuir metadados de qualidade aos dados**.

As exigências sobre o nível de descrição e de atribuição de metadados devem ser identificadas desde o começo do seu projeto e revistas ao longo do ciclo de vida dos seus dados. **Essa é a essência de uma boa curadoria de dados**. Mas lembre-se de que o esforço que você vai investir em documentar os seus dados depende da vida útil e do nível de compartilhamento que você pretende para eles.

Uma estratégia interessante é **descrever primeiramente o seu projeto de pesquisa**, isto ajudará na contextualização da pesquisa e justificará a razão dos dados que ela precisará coletar; e **depois descrever os dados** propriamente ditos com todas as suas especificidades técnicas e estruturais.

Bom lembrar que quando um protocolo de pesquisa é utilizado, grande parte da documentação necessária já existe. Se instrumentação é utilizada, calibração e outros ajustes necessários para captura dos dados devem ser registrados; outras informações importantes devem ser registradas, como parâmetros, unidades, valores codificados etc..

O **caderno de laboratório** talvez seja a mais rigorosa forma de documentação de sua pesquisa, portanto, considere incluir uma versão digital dele na documentação que acompanha os dados.

A **documentação viabiliza a compreensão e transmissão para o futuro dos significados e conhecimentos que os dados portam.** Portanto, a recomendação mais importante é que você descreva os seus dados tomando como referência um usuário que não está familiarizado com seu projeto de pesquisa, com o ambiente de pesquisa, e a metodologia adotada. Além do mais, como os seus dados vão ser arquivados para uso futuro, a documentação deve ser escrita para instruir usuários que **estarão num horizonte temporal de 20 ou mais anos adiante**.

5.1
DOCUMENTAÇÃO QUE DESCREVE OS DADOS

Além dos metadados é importante que o pesquisador deposite **arquivos adicionais que descrevam as suas coleções de dados com mais detalhes**, especialmente os processos usados para criá-las. Sem essa documentação uma coleção de dados pode não estar em condições de ser reusada.

QUE INFORMAÇÕES DEVEM ESTAR PRESENTE NA DOCUMENTAÇÃO DOS DADOS?

A documentação que deve acompanhar os dados PODE ser apresentada de duas formas:

- **METADADOS**
  Conjunto de informações essenciais, padronizadas e estruturadas que documentam os dados explicando sobre a proveniência, origem, propósito, autores, instituições envolvidas, termos de uso e detalhes técnicos e estruturais.

- **DOCUMENTOS**
  Consiste de documentos em formatos variados – texto, vídeos, planilhas, etc. – que apoiam o entendimento sobre como os dados foram coletados, gerados, processados e de como estão estruturados, organizados, nomeados. Podem incluir, por exemplo: cadernos de laboratório e caderno de campo, guia de usuário, questionários, lista de parâmetros, *codebook* para dados estatísticos, especificação de formatos e descrições textuais.

5.2
METADADOS

QUAL O PAPEL DOS METADADOS NA DESCRIÇÃO DOS DADOS?

Uma parte da documentação que acompanha os dados é expressa por meio de **METADADOS**, que de uma forma simples são "dados sobre dados". Os metadados são formados por conjuntos de etiquetas ou campos definidos de forma padronizada, que são coletivamente chamados de esquema ou formato de metadados. Os metadados identificam informações importantes sobre os dados, por exemplo, o metadado cuja etiqueta é "AUTOR" informa quem são os autores dos dados.

Os metadados servem como base para buscas mais estruturadas e consistentes em base de dados e repositórios de dados, facilitando a **descoberta das coleções de dados** pela comunidade científica e pelo público em geral.

Os metadados podem ser **usados por pessoas e por programas de computador** para ajudar a **descobrir, integrar e analisar dados**.

Assinalar metadados detalhados também **protege o investimento na geração dos dados**. Mudanças na tecnologia, equipes ou mesmo o efeito do tempo na memória das pessoas pode causar perdas de informação. Manter registros na forma de metadados sobre os

dados protege-os contra perdas de detalhes importantes, assegurando a usabilidade dos dados ao longo do tempo[31].

Para um dado projeto de pesquisa que envolva a coleta e/ou geração de dados de pesquisa, metadados são geralmente criados em dois níveis:

- **NÍVEL DE PROJETO**
  Descreve o projeto de pesquisa, estabelecendo o contexto para a compreensão da razão da coleta/geração de dados e como eles serão usados.

- **NÍVEL DE DADOS**
  Descreve os dados e as coleções de dados com ênfase nos detalhes técnicos.

QUE INFORMAÇÕES BÁSICAS SOBRE O PROJETO EU DEVO REGISTRAR?

Uma forma interessante de identificar quais informações você deve registrar sobre o seu projeto de pesquisa é pensar os metadados em termos de **POR QUE, QUEM, O QUE, QUANDO e ONDE[32].** Embora a estrutura da documentação dos dados possa tomar outra forma, responder essas questões vai ajudar você a assegurar uma descrição completa e um contexto importante para os dados, particularmente, no decorrer do tempo.

- **POR QUE**: **objetivo/justificativa/relevância** do projeto (resumo).
- **QUEM**: **equipe envolvida** com o projeto (líder, pesquisadores, técnicos, etc.).
- **ONDE**: localização e descrição dos **ambientes estudados**.
- **QUANDO**: **intervalo de tempo** considerado pelo projeto.
- **COMO**: descrição da **metodologia** do projeto

[31] DATAONE. **Tutorials on data management Lesson 7**: Metadata. Disponível em: <https://www.dataone.org/sites/all/documents/L07_Metadata.pptx>. Acesso em: 01 out. 2015.
[32] MICHENER, William K. et al Nongeospatial metadata for the ecological sciences. **Ecological Appliations**, v.7, n.1, p. 330-342, 1977. Disponível em: <http://lits.bio.ic.ac.uk:8080/litsproject/Micheneretal1997.pdf>. Acesso em: 10 out. 2015.

Exemplo de metadados de PROJETO	
• NOME DO PROJETO • DESCRIÇÃO DO PROJETO • LIDER DO PROJETO • PESQUISADORES • INSTITUIÇÕES ENVOLVIDAS	• AMBIENTES DE PESQUISA • DURAÇÃO DO PROJETO • FINANCIADOR DO PROJETO • PROJETO GUADA-CHUVA • CONTATO PARA INFORMAÇÕES • ASSUNTO/PALAVRAS-CHAVE

QUE INFORMAÇÕES BÁSICAS SOBRE OS DADOS EU DEVO REGISTRAR?[33, 34]

- **POR QUE os dados foram coletados?**
  Descreve o contexto científico da criação dos dados: questão de pesquisa; propósito cientifico da coleta de dados; que dados foram coletados e um breve resumo da coleção de dados.

- **QUEM coletou os dados?**
  Descreve as pessoas envolvidas e os *stakeholders*:
  Quem coletou os dados e **quem financiou**; quem contatar para **mais informações** sobre os dados; **como citar os dados** de forma que as pessoas envolvidas tenham o devido crédito.

- **O QUE os dados incluem?**
  Para descrever os dados, várias categorias de detalhes são necessárias, por exemplo:

  - **Contexto digital**: nome da coleção de dados; nomes dos arquivos que compõem a coleção; formato dos arquivos; data das modificações; lista de coleções de dados relacionadas e ensilares; *software* (incluindo o número da versão) usado para preparar e ler a coleção de dados; procedimentos de processamento de dados.

[33] WIGGINS, Andrea et al. **Data management guide for public participation in scientific research**. Albuquerque, NM: DataONE, 2013. Disponível em: <https://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf>. Acesso em: 10 out. 2015.

[34] STRASSER, Carly et al. **Primer on data management**: what you always wanted to know. California: CDL, Feb. 2012. Disponível em: <https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf>. Acesso em: 01 out. 2015.

- **Detalhes sobre Parâmetros/Variáveis**: como cada um foi medido ou produzido; unidades de medida, formatos dos dados, precisão, acurácia e incerteza.

- **Informações sobre os dados**: taxonomias, codificação, procedimentos de controle e garantia de qualidade, bias de amostragem entre outros.

- **Conteúdo dos arquivos de dados:** definição dos parâmetros e explicação sobre seus formatos, notas de revisão de qualidade, valores faltantes.

- **Informações complementares sobre os dados:** usando taxonomias padronizadas quando possível.

- **Organização dos dados**: relacionamento entre as entidades de dados, arquivos, diretórios, e/ou tabelas do banco de dados; quando possível, incluir o diagrama da estrutura do banco de dados.

- **QUANDO os dados foram coletados?**

  A extensão temporal e a resolução dos dados devem ser o mais específico possível, registrando ano, mês, dia, tempo da forma mais apropriada aos objetivos do projeto. Três facetas de temporalidade devem ser incluídas na descrição dos dados juntamente com os formatos dos dados.

  - **Limites temporais**: a faixa de tempo total de **observação** incluída na coleção de dados
  - **Extensão temporal da coleção de dados**: a faixa total de **aquisição** de dados.
  - **Resolução temporal**: descreve a frequência na qual os dados são coletados ou adquiridos.

- **ONDE os dados foram coletados?**

  Assim como os aspectos temporais, informações sobre a localização são de grande importância para o uso e reuso dos dados. Três características devem ser consideradas:

  - **Extensão espacial:** descreve os limites geográficos da coleção de dados
  - **Resolução espacial**: descreve a especificidade de espaçamento para a localização.
  - **Formatos de dados espaciais**: descreve os formatos de dados espaciais.

- **COMO os dados foram coletados?**

  Esta é uma questão chave para a interpretação e reuso dos dados, pois descreve as metodologias usadas na coleta dos dados: protocolos da coleta de dados, técnicas de mensuração, métodos de garantia e controle de qualidade para todas as coleções; instrumentos de medida usados (incluindo modelo e número de série); padrões e calibração utilizados.

Exemplo de metadados das COLEÇÕES DE DADOS	
• TÍTULO DA COLEÇÃO • IDENTIFICADOR PERSISTENTE DA COLEÇÃO • RESUMO DA COLEÇÃO • AUTORES • CITAÇÃO DA COLEÇÃO • DATA DA PUBLICAÇÃO DOS DADOS • PERÍODO DE TEMPO QUE COBRE • ESCOPO GEOGRÁFICO • DIREITOS ASSOCIADOS AOS DADOS • ASSUNTO/PALAVRAS-CHAVE	• ORIGEM DOS DADOS (experimental, observacional, crus, derivados...) • TIPO DE DADOS (inteiro, imagem...) • INSTRUMENTOS USADOS • AQUISIÇÃO DOS DADOS • TIPO DE ARQUIVO • MÉTODO DE PROCESSAMENTO DE DADOS (software) • PROCESSAMENTO DE DADOS (scripts ou código) • PARÂMETROS DA COLEÇÃO DE DADOS • CONTROLE DE QUALIDADE

O QUE É ESQUEMA (OU FORMATO) DE METADADOS?

O conjunto de unidades de informação – por exemplo, “titulo”, “autor”, “resumo” - necessário para a descrição de um recurso como um livro ou uma coleção de dados – é chamado de **esquema ou formato de metadados**.

Vários esquemas de metadados foram desenvolvidos e muito deles são aplicáveis na descrição de dados de pesquisa. Há os esquemas gerais, aplicáveis a várias categorias de informação e há os esquemas voltados para disciplinas específicas, como Ciências Ambientais. Muitos desses esquemas são padrões de fato e de direito e contribuem também para a troca de informações (interoperabilidade) entre repositórios e base de dados.

Solicite auxílio aos profissionais de informação de sua instituição para saber qual o esquema de metadados mais conveniente para descrever os seus dados.

Observe também quais são os **padrões de metadados utilizados pelos repositórios** ou centros de dados onde você planejou, no seu PGD, arquivar os seus dados por longo prazo. Dessa forma você economizará tempo assinalando os metadados corretos.

COMO E ONDE EU REGISTRO OS METADADOS QUE DESCREVEM MEUS DADOS?

Os pesquisadores geralmente criam registros de metadados para os seus dados preenchendo formulários eletrônicos ou editores de metadados voltados para depósitos de dados de banco de dados, repositórios digitais ou centros de dados.

Caso sua instituição não disponha de repositório de dados ou outro dispositivo para arquivamento dos dados e você não tenha ainda a aceitação de um centro de dados, use uma planilha para registrar os metadados e junte com a documentação em uma pasta, não esquecendo de fazer *backups* desses arquivos.

O PAPEL DO ESQUEMA DUBLIN CORE

A maioria dos repositórios, para cada coleção de dados depositada, inclui campos de metadados complementares, que estão em conformidade com algum formato ou esquema mais geral que facilite a busca e a troca de metadados (interoperabilidade) entre repositórios. O Dublin Core (DC)[35], que é um esquema simples para descrição de recursos da internet, cumpre esse papel. O DC inclui campos descritivos para autor, resumo, fonte, assuntos, formatos, perfazendo um total de 15 campos.
No âmbito de um repositório, o esquema DC pode ser configurado segundo um padrão baseado em XML, conhecido como Protocolo OAI-PMH[36], que permite que os metadados possam ser coletados automaticamente por outros sistemas chamados de "provedores de serviço".

5.3
MAIS SOBRE DESCRIÇÃO DE DADOS DE PESQUISA:

a) IDENTIFICAÇÃO

COMO IDENTIFICAR AS MINHAS COLEÇÕES DE DADOS?

- **IDENTIFICADORES PERSISTENTES**
  Na qualidade de objetos digitais, as coleções de dados devem ser identificadas por esquemas de **identificação persistentes, padronizados e globalmente únicos**. A identificação apropriada vai permitir que seus dados sejam preservados, recuperados, citados e compartilhados.

[35] <http://dublincore.org/>
[36] <https://www.openarchives.org/pmh/>

- **TÍTULOS DESCRITIVOS**

  Títulos descritivos, ou seja, que portem informações sobre as características dos dados, também são importantes para que eles sejam identificados por outros pesquisadores que procuram por eles.

ATRIBUA IDENTIFICADORES PERSISTENTES PARA OS CONJUNTOS DE DADOS

Escolha um esquema de identificação padronizado e de amplo uso para o tipo de dados e para a área específica que está sendo considerada.

O **DOI** (Digital Object Identifier)[37] é amplamente usado para artigos de periódicos científicos e se aplica também a coleções de dados. Outros esquemas de identificação importantes para a área científica são:

- **URI** (Uniform Resource Identifier)[38]
- **PURL** (Persistent Uniform Resource Locator) [39]
- **HDL** (The Handle System) [40]
- **InChi** (IUPAC International Chemical Identifier) [41]

ATRIBUA TITULOS DESCRITIVOS PARA AS SUAS COLEÇÕES DE DADOS

A recomendação é que os títulos das coleções de dados sejam os mais descritivos possíveis. Lembre-se de que o título é frequentemente a primeira coisa que um interessado verá quando estiver procurando por uma *data set*, dessa forma, atribuir títulos com significado facilita a vida de quem está procurando por eles.

É importante assinalar que os dados serão acessados no futuro por **pessoas que desconhecem os detalhes do projeto** que gerou os dados, isso torna ainda mais relevante um título representativo que possa ajudar a interpretação dos dados daqui a alguns anos.

[37] http://www.doi.org
[38] http://www.ietf.org/rfc/rfc2396.txt
[39] http://www.purl.org/
[40] http://www.handle.net/
[41] http://www.iupac.org/inchi

5.3
MAIS SOBRE DESCRIÇÃO DE DADOS DE PESQUISA:

b) ORGANIZAÇÃO DO CONJUNTO DE DADOS[42]

DESCREVA A ORGANIZAÇÃO GLOBAL DO SEU CONJUNTO DE DADOS

Comumente um conjunto ou coleção de dados contém um **grande número de arquivos que são relacionados.** Esses arquivos estão organizados em pastas, **diretórios ou mesmo em tabelas de um banco de dados**. A descrição e documentação da organização adotada se tornam, portanto, críticas para quem quer **localizar e usar** os seus dados, incluindo você mesmo e os seus colegas de pesquisa.

DOCUMENTE A RELAÇÃO ENTRE ARQUIVOS E PASTAS OU ENTRE AS TABELAS DO BANCO DE DADOS

Minimamente, a organização e o relacionamento entre diretórios, arquivos ou tabelas de um banco de dados e ainda outros materiais de apoio precisam ser completamente descritos. Use o resumo (*abstract*) que descreve os seus dados para documentar também como eles estão organizados: que tabelas contêm, onde estão localizados os materiais de apoio, os metadados, ou ainda outros documentos relevantes.

O **relacionamento entre entidades de dados** deve ser descrito e documentado para possibilitar a **compreensão pelos futuros usuários** e pelos repositórios que hospedam os dados; portanto, considere representar o relacionamento lógico entre as entidades de dados usando um **Diagrama Entidade Relacionamento** proveniente do MER (Modelo Entidade Relacionamento).

DESCREVA O CONTEÚDO DOS SEUS ARQUIVOS DE DADOS

Para que outros pesquisadores possam utilizar os dados que você coletou/gerou, eles devem compreender integralmente o conteúdo dos conjuntos de dados. Para isso, a documentação que deve acompanhar os dados precisa apresentar uma descrição completa dos **PARÂMETROS**, incluindo os **NOMES DOS PARÂMETROS**, como eles foram **MEDIDOS**, as **UNIDADES DE MEDIDAS**, as **ABREVIATURAS**, os **FORMATOS**, e a definição dos **VALORES CODIFICADOS**[43].

A descrição dos dados deve ser acompanhada de arquivos de dados do tipo "**readme.txt**", um arquivo de metadados usando um esquema padronizado, ou ambos.

---

[42] DATAONE. **Tutorials on data management Lesson 7:** Metadata. Disponível em: <https://www.dataone.org/sites/all/documents/L07_Metadata.pptx>. Acesso em: 01 out. 2015.

[43] HOOK, Les A. et al. **Best Practices for Preparing Environmental Data Sets to Share and Archive**. Oak Ridge: Oak Ridge National Laboratory, September 2010. Disponível em: <http://daac.ornl.gov/PI/BestPractices-2010.pdf>. Acesso em: 10 out. 2015.

- **NOME DOS PARÂMETROS**

  Os parâmetros reportados no *conjunto de dados* precisam ter nomes que descrevam claramente o conteúdo que eles denotam; é importante que os **nomes** estejam **padronizados no âmbito geral do projeto de pesquisa**. Use preferencialmente **nomes já consagrados** na área, por exemplo, "Temp" para temperatura. Procure ser consistente na grafia dos nomes, por exemplo, na capitalização (temp, Temp, TEMP).

- **UNIDADES**

  É muito importante que as unidades sejam definidas de forma que outros pesquisadores compreendam o que está sendo reportado, portanto as **unidades que dimensionam os parâmetros necessitam estar explicitamente estabelecidas** no arquivo de dados e na documentação. Recomenda-se as unidades preconizadas pelo **SI** (**Sistema Internacional de Unidades**)[44] quando as especificidades das disciplinas não forem um impedimento.

A recomendação é que **não use abreviaturas** quando você estiver descrevendo as unidades. Por exemplo, a unidade para respiração é: *moles de dióxido de carbono por metro quadrado por ano.*

- **FORMATOS DOS PARÂMETROS**

  Para cada conjunto de dados, você deve escolher um **formato para cada parâmetro**, não se esquecendo de explicar os formatos selecionados na documentação. Use de forma consistente esses formatos para toda a coleção de dados. Formatos consistentes são particularmente críticos para DATAS, TEMPO e COORDENADAS ESPACIAIS. Por exemplo: use para datas: yyyy-mm-dd ou dd-mm-yyy; para tempo use a notação 24 horas, registrando o horário, local e o tempo UTC (Tempo Universal Coordenado).

- **CAMPOS CODIFICADOS**

  Campos codificados geralmente são preenchidos tendo como base **listas padronizadas**, como por exemplo, uma lista de siglas de instituições ou a representação dos elementos da tabela periódica. Uma grande vantagem dos campos codificados é que eles são **mais eficientes para armazenamento e recuperação** de dados do que os campos de texto livre. Nessa direção, você pode estabelecer seus próprios campos codificados, definindo valores para serem usados de forma consistente em vários arquivos de dados. Um lembrete importante é que você deve estar atento para as mudanças que possam ocorrer nos esquemas de códigos, principalmente os esquemas definidos por agentes externos; essas **mudanças devem ser documentadas**.

- **VALORES AUSENTES**

  É importante usar de forma consistente a **notação de valores ausentes** para campos numéricos e textuais do arquivo de dados. Nessa direção, um **valor codificado para os valores ausentes** deve ser definido.

  A forma preferencial de identificar um dado ausentes é por meio de um **campo vazio** (NULL= sem valor);

[44] http://pt.wikipedia.org/wiki/SI

- Se por alguma razão não for possível deixar uma célula vazia, então use um valor extremo (por exemplo, -9999) em campos numéricos;
- Para campos textuais use NA (“Não se Aplica”) ou ND (“Não Disponível”);
- Use *data flags* em uma coluna separada de uma planilha para qualificar as células vazias. Por exemplo: “M1= ausente; amostra não coletada”

5.4
COMO GARANTIR QUE A DOCUMENTAÇÃO SEJA LIDA NO FUTURO?

A documentação também precisa ser lida ao longo do tempo, para que as coleções de dados possam ser compreendidas e contextualizadas pelo usuário em algum momento no futuro, portanto ela precisa seguir procedimentos que garantam que seu conteúdo possa ser interpretado no longo prazo.

COMO GARANTIR QUE A DOCUMENTAÇÃO SEJA LIDA NO FUTURO?

- **FORMATOS ESTÁVEIS**
  É necessário assegurar que a documentação esteja num formato estável, não proprietário e independente de *software*. Por exemplo, se fotografias, mapas, equações ou desenhos precisam ser incluídos, use um formato não proprietário como HTML; use o formato de arquivo JPG para incluir as imagens individuais e formatos MP4 para vídeos (veja o capítulo 9: Formate os seus dados).

- **CONVERSÃO DE FORMATOS**
  Converter documentos textuais mais elaborados para um formato estável como o PDF ou **PDF/A** – que é uma norma ISO - é uma opção que deve ser considerada.

- **ARQUIVO “LEIAME.TXT”**
  A documentação deve estar em arquivos separados dos arquivos de dados; crie uma pasta específica para hospedá-la e um arquivo “leiame.txt” para explicar o seu conteúdo. Esse arquivo pode ser de grande valia no futuro.

,

6
PROTEJA SEUS DADOS
BACKUP, ARQUIVAMENTO E PRESERVAÇÃO[45]

Os termos SEGURANÇA DE DADOS, *BACKUP* DE DADOS, ARQUIVAMENTO DE DADOS e PRESERVAÇÃO DE DADOS são utilizados frequentemente como sinônimos. Porém, é importante enfatizar que eles têm significados e propósitos diferentes, especialmente quando se trata da gestão de dados de pesquisa. Vejamos com um grau a mais de precisão o que denotam esses termos:

- **SEGURANÇA DE DADOS**
  E o termo mais amplo, pois cobre uma ampla variedade de tópicos, incluindo *backup*, arquivamento, preservação e proteção física, criptografia e ainda as leis que governam a proteção dos dados.

Os termos "*backup*" e "arquivamento de dados", apesar de estarem relacionados ao salvamento de uma versão específica de um arquivo, eles são processos bem distintos e ocorrem em momentos diferentes do ciclo de gestão de dados.

- **BACKUP**
  O termo é usado especificamente quando se faz várias cópias de vários arquivos tendo conhecimento que os arquivos podem mudar. Dessa forma, as cópias de *backup* podem ser guardadas por certo período de tempo, mas podem ser descartadas quando for conveniente.

  Quando um pesquisador faz o *backup* de um arquivo de dados, ele está tirando um retrato (ou uma cópia) dos dados naquele preciso momento; esta cópia será usada para restaurar a versão original caso ela tenha sido, por algum motivo, perdida, corrompida, destruída, ou alterada.

  Por sua condição transitória, os *backups* são armazenados por prazos determinados, curtos ou médios, que dependem das necessidades do usuário e dos procedimentos da instituição. Além do mais, eles são efetuados regularmente de acordo com um cronograma pré-estabelecido.

[45]DATAONE. **Tutorials on data management Lesson 06:** Protecting Your Data: Backups, Archives, and Data Preservation. Disponível em: <https://www.dataone.org/sites/all/documents/L06_DataProtectionBackups.pptx >. Acesso em: 01 out. 2015.

- **ARQUIVAMENTO**
  É usado quando um arquivo deve ser preservado como está, ou quando se deseja um registro do histórico do arquivo. Geralmente, o arquivamento lida com registros que estão na sua versão final e faz parte das etapas requeridas para a preservação dos dados para necessidades futuras, ou seja, para a preservação de longo prazo. Nessa direção, o arquivamento é realizado, via de regra, quando o projeto termina.

- **PRESERVAÇÃO**
  O termo engloba muitas das metodologias utilizadas pelos processos de *backup* e de arquivamento, entretanto inclui outros itens, tais como: resgate de dados, reformatação de arquivos, conversão de dados e atribuição de metadados.

Neste capitulo serão colocadas as recomendações relativas aos processos de BACKUP e PROTEÇÃO FISICA. As questões sobre ARQUIVAMENTO e PRESERVAÇÃO, pela importância do tema, serão tratadas especificamente no capítulo seguinte.

6.1 BACKUP

Fazer *backup* dos seus dados e mantê-lo atualizado é uma etapa essencial da gestão de dados de pesquisa. **Backups regulares protegem seus dados contra perdas acidentais e intencionais** e podem ser usados para restaurar os dados originais evitando a perda definitiva dos dados.

Perdas acidentais ou intencionais de dados podem ser causadas por:

- **FALHA DE HARDWARE, SOFTWARE OU MIDIA;**
- **INFECÇÃO POR VIRUS OU ATAQUE DE HACKERS;**
- **FALHA DE ENERGIA;**
- **ERRO HUMANO CAUSANDO DELEÇÃO OU MUDANÇA NOS ARQUIVOS.**

A escolha do procedimento de *backup* que deve ser adotado vai depender das circunstâncias locais, o valor percebido dos dados e do **nível de risco considerado aceitável para os dados** considerados. Para muitos pesquisadores, realizar uma análise de risco informal pode fornecer uma boa indicação para as necessidades de *backup* para os dados.

Porém, no âmbito de sua instituição, departamento ou laboratório, estabelecer uma política de *backup* para os dados, considerando a importância que eles têm hoje para as atividades de pesquisa, torna-se algo importante; além do mais, a padronização facilita os procedimentos de segurança.

COMO FAZER *BACKUP* DOS MEUS DADOS?

- **MANUALMENTE**
  Se você precisa fazer *backup* de somente uns poucos arquivos, isto pode ser realizado sem dificuldades manualmente; isso implica, porém, **lembrar** de fazer os *backups* na regularidade necessária.

- **AUTOMATICAMENTE**
  Caso você tenha muitos arquivos, ou não quer ficar lembrando de fazer *backups*, você pode utilizar um *software* que faça isso automaticamente. Muitos computadores já têm *software* próprio de *backup*, assim como os drives de disco externo.

Uma regra de boas práticas é **não fazer, se possível, *backups* manualmente**. Os sistemas **automáticos farão o serviço melhor e mais rápido**.

Se houver um **suporte de TI**, acione-o para **ajudá-lo com os seus *backups***, mas não assuma que alguém fará os *backups* por você. Mesmo que alguém o faça, assegure-se de que os *backups* foram **plenamente testados**.

DEVO FAZER BACKUP DE UM ARQUIVO DE DADOS ESPECÍFICO OU DE TODO O SISTEMA?

O que você precisa restaurar caso haja um evento de perda de dados? Se a sua instituição pode restaurar todo o sistema, então você pode se responsabilizar somente pelos seus arquivos de dados; caso contrário, você tem que se responsabilizar pelos *backups* do sistema necessários, por exemplo, pela visualização dos dados.

COM QUE FREQUÊNCIA DEVO FAZER BACKUP DOS MEUS DADOS? CONTINUAMENTE? DIARIAMENTE? SEMANALMENTE? MENSALMENTE?

Para reduzir os riscos aos menores níveis possíveis, uma boa regra é fazer *backup* **a cada alteração que você fizer nos dados, ou em intervalos regulares**.

Use processos automáticos de *backup* para os arquivos de dados usados frequentemente e para os arquivos críticos.

Considere as seguintes questões na sua análise de criticidade dos arquivos de dados:

- **Você pode se permitir perder semanas de coleta de dados caso você faça *backup* somente uma vez por mês?** Se a resposta for não, você deve considerar fazer cópias de segurança mais frequentemente.

- **Você está criando dados em tempo real que não podem ser reproduzidos?** Se a resposta for sim, você deve fazer *backups* continuamente.

- Considere também o **custo-benefício em termos de criticidade e importância dos dados**; considere ainda a infraestrutura de *software* e de *hardware* necessários para rodar o sistema de *backup*.

Você deve atentar também para o **tempo de retenção** do seu *backup*. Uma boa prática é manter o *backup* corrente localmente, **mantendo os três *backups* prévios *off-site***, fazendo a rotação quando uma nova cópia é produzida.

QUE TIPO DE *BACKUP* DEVO FAZER?

Existem dois tipos de *backups*: *backup* completo e *backup* parcial ou incremental.

- ***BACKUP* PARCIAL OU INCREMENTAL**
  Faz cópia de segurança apenas do que foi mudado desde o último *backup*. Dado que você está fazendo *backup* de somente uma parte do seu sistema, é mais fácil, mais rápido e requer menos recursos em termos de processamento e de espaço de armazenamento.

- ***BACKUP* COMPLETO**
  Faz cópia de segurança de todo os seus dados. Inicialmente você deve fazer o *backup* completo e, nas próximas operações, você pode fazer o *backup* parcial que fará cópia de todos os dados que sofreram alguma mudança desde o último *backup*.

ONDE DEVO ARMAZENAR O *BACKUP* DOS MEUS DADOS?

A sua instituição ou o seu projeto devem ter um lugar específico onde as cópias de segurança serão armazenadas, verifique isso com a equipe de TI, pois é mais conveniente manter os *backups* em **unidades de discos em rede**.

Se a sua instituição não dispõe de um sistema de *backup*, você pode considerar o uso de discos externos, fita magnética ou armazenamento *online* ou usar serviços de armazenamento nas nuvens como os oferecidos pelo Dropbox[46], Amazon[47] ou pelo Google[48].

Estas opções podem depender da quantidade e do tipo de arquivo de dados que você precisa proteger. Se você está fazendo *backup* diários de pequenos arquivos, provavelmente DVD/BlueRay gravável pode ser suficiente; mas se lidando com grandes

[46] <https://www.dropbox.com/>
[47] <https://aws.amazon.com/pt/>
[48] <https://cloud.google.com/storage/>

volumes de dados, é mais conveniente usar discos externos ou fita magnética (fita padrão LTO Ultrium)[49].

Como regra geral, é recomendável fazer múltiplas versões dos seus *backups*, assegurando **que eles estejam armazenados em diferentes tipos de mídias e de formatos**. Lembrando que para assegurar o acesso futuro aos arquivos use preferencialmente formatos padronizados e não proprietários

ARMAZENAMENTO *OFF-SITE*

Mesmo que você já tenha um *backup* no seu local de trabalho, **é desejável que você tenha uma cópia de segurança em outro local, preferencialmente em outro prédio**. Isso evita que ambas as versões dos seus dados – original e *backup* – sejam destruídas em caso de algum sinistro ou outro incidente em seu escritório, laboratório ou em sua casa, isso é especialmente importante para os dados críticos e de difícil obtenção.

O QUE É POLÍTICA DE *BACKUP* E PARA QUE SERVE?

É uma boa prática criar um **documento que estabelece todas as orientações, procedimentos e responsabilidades acerca das cópias de segurança dos dados** de pesquisa no contexto do seu projeto ou laboratório. O documento deve ser revisado periodicamente, posto que hardware, software, projetos, equipes estão sempre mudando. Ele deve conter:

,

- PAPÉIS;
- RESPONSABILIDADES;
- ONDE SERÃO ARMAZENADOS OS *BACKUPS;*
- COM QUE FREQUÊNCIA OS *BACKUPS* SERÃO REALIZADOS;
- COMO ACESSAR OS ARQUIVOS DO *BACKUP*; COMO RESTAURAR OS DADOS;
- FORMATOS DE ARQUIVO RECOMENDADOS;
- PROCEDIMENTOS PARA MIGRAÇÃO DOS DADOS, PARA ASSEGURAR QUE OS DADOS NÃO SEJAM PERDIDOS POR DEGRADAÇÃO DAS MÍDIAS OU MUDANÇA NOS FORMATOS.

[49] <https://pt.wikipedia.org/wiki/Linear_Tape-Open >

6.2
SEGURANÇA DOS DADOS[50]

SEGURANÇA FÍSICA, SEGURANÇA DE REDE E SEGURANÇA DO COMPUTADOR E DE ARQUIVO

Segurança física, segurança de rede e segurança do computador e dos arquivos precisam ser consideradas para garantir a proteção dos dados e prevenir acessos não autorizados, alterações, divulgação inapropriada ou destruição desses recursos informacionais.

Entretanto, as configurações de segurança de dados **precisam ser proporcionais à natureza dos dados e do risco envolvido**.

Atenção! A segurança dos dados é importante também no **momento em que os dados precisam ser destruídos**.

A segurança dos dados pode também ser necessária para **proteger os direitos de propriedade intelectual, interesses comerciais** – por exemplo, dados que serão usados para patenteamento - ou para **manter sigilo sobre dados pessoais ou para proteger informações sensíveis**.

O QUE É NECESSÁRIO PARA GARANTIR A SEGURANÇA FÍSICA DOS DADOS?

- **ACESSO FÍSICO**
  Controlar o acesso a salas e edifícios onde os dados, computadores e mídias são mantidos.

- **REGISTROS DE EVENTOS (*LOGGING*)**
  Manter registro da remoção de ou acesso a mídias ou cópias impressas na área de armazenamento.

- **TRANSPORTE**
  Transportar dados sensíveis apenas em circunstâncias excepcionais, mesmo quando é necessário reparar algum equipamento. Por exemplo, entregar um disco rígido contendo dados sensíveis a um fabricante ou técnico para manutenção, pode causar uma brecha importante na segurança dos dados.

[50] EYNDEN, Veerle et al. **Managing and data sharing:** best practice for researchers. Colchester: UK Data Archive, 2011. Disponível em: <http://www.data-archive.ac.uk/media/2894/managingsharing.pdf >. Acesso em: 01 out. 2015.

O QUE SIGNIFICA SEGURANÇA DE REDE?

- **ARMAZENAMENTO DE DADOS CONFIDENCIAIS**
  Não armazenar dados confidenciais – por exemplo, dados contendo informações pessoais – em servidores ou em outros computadores conectados a redes externas, particularmente em servidores que hospedam serviços internet.

- **FIREWALL**
  Utilizar proteção do *firewall* é a segurança proporcionada pelas atualizações e *patches* do sistema operacional, no intuito de evitar vírus e códigos maliciosos.

O QUE SIGNIFICA SEGURANÇA DO COMPUTADOR E DOS ARQUIVOS?

- **BLOQUEIO DO COMPUTADOR**
  Bloquei o seu computador com uma senha e instale um **sistema de *firewall*.**

- **OSCILAÇÃO DE ENERGIA**
  Proteja seu computador contra a oscilação de energia instalando uma **fonte de alimentação secundária ininterrupta** *off-line* (line-interactive), também conhecida como UPS (*uninterruptible power supply*) ou *no-break*.

- **SENHAS PARA OS ARQUIVOS**
  Implemente proteção por **senhas e controle de acesso aos arquivos de dado**.

- **ACORDO DE NÃO DIVULGAÇÃO**
  Estabeleça um **Acordo de Não Divulgação** (*Non-Disclosure Agreement* ou NDA em inglês) para os gestores ou usuários dos **dados confidenciais**.

- **ENVIO DE ARQUIVOS**
  Não envie dados pessoais ou confidenciais via *e-mail* ou por FTP (Protocolo de Transferência de Arquivo). Esses dados devem ser criptografados antes de ser enviados.

- **DESTRUIÇÃO DOS DADOS**
  Destrua os dados, quando necessário, de forma consistente, atente para as normas voltadas para essa questão. **Apagar os arquivos ou formatar os discos não são procedimentos seguros**.

6.2.1
SEGURANÇA DE DADOS PESSOAIS

Dados que contém informações pessoais devem ser tratados **com um alto nível de segurança** que vai muito além dos dados que não tratam desse tipo de informação.

É importante observar que dados pessoais **podem existir também em formatos não digitais**, por exemplo, como registros de pacientes, formulários de consentimento assinados ou folha de rosto de entrevistas. Estes itens devem ser protegidos da mesma forma que os arquivos digitais.

QUE AÇÕES PODEM FACILITAR A PROTEÇÃO DE DADOS PESSOAIS?

- **FAÇA ANONIMIZAÇÃO OU AGREGAÇÃO DOS DADOS.**
- **SEPARE OS DADOS DE ACORDO COM AS NECESSIDADES DE SEGURANÇA.**
- **REMOVA INFORMAÇÕES PESSOAIS**
  Informações tais como nomes e endereços devem ser removidas dos arquivos de dados e armazenadas separadamente.
- **CRIPTOGRAFE OS DADOS CONTENDO INFORMAÇÕES PESSOAIS**
  A criptografia é essencial para os dados que vão ser transmitidos. Processos de encriptação devem também ser executados antes dos dados serem armazenados.

A forma como os dados confidenciais e os dados contendo informações pessoais serão armazenados deve ser negociada durante a fase de estabelecimento do **consentimento informado**. Isto assegura que as pessoas a quem pertencem os dados foram informadas e concordaram com a forma como eles serão armazenados e transmitidos.

6.2.2
TRANSFERÊNCIA DE ARQUIVOS E CRIPTOGRAFIA

Transmitir dados entre diferentes locais ou internamente entre os membros de sua equipe de pesquisa pode ser algo desafiador para a infraestrutura de gestão de dados. **Para garantir que dados pessoais e sensíveis possam ser seguramente transmitidos eles devem ser criptografados segundo um padrão seguro**. Somente os dados que passaram comprovadamente por processos de anonimização podem ser transmitidos sem estarem criptografados.

A criptografia mantém a segurança dos dados durante a transmissão.

POSSO CONFIAR NO CORREIO ELETRÔNICO PARA TRANSMITIR MEUS DADOS? Confiar no e-mail para transferir coleções de dados, mesmo internamente, se configura como um ponto vulnerável na proteção de dados sensíveis. Qualquer coisa enviada por e-mail circula e permanece por muitos servidores, portanto devem ser criptografados segundo padrões apropriados.
COMO POSSO TRANSMITIR COM SEGURANÇA GRANDES ARQUIVOS DE DADOS? A pesquisa cientifica produz, em larga escala coleções de dados, que muitas vezes estão na forma de arquivos volumosos. Transferir esses dados pode ser problemático. Serviços comerciais de compartilhamento existem para facilitar o movimento de arquivos, entretanto alguns serviços não são necessariamente permanentes e seguros e frequentemente estão localizados no exterior e não são cobertos pela legislação do país. **Se gerenciado e controlado por instituições responsáveis, um serviço de DROPBOX pode ser uma solução segura para a transferência de grandes arquivos.** Destaca-se que a necessidade de criptografia para arquivos com dados pessoais e sensíveis antes da submissão ao serviço permanece.

6.2.3
ELIMINAÇÃO DE DADOS[51]

Ao longo do processo de pesquisa, cópias de arquivos de dados que **não são mais necessários precisam ser destruídas.** Quando a pesquisa é concluída, **arquivos de dados que não serão preservados precisam ser eliminados de forma segura após a conclusão da pesquisa.**

Estratégias confiáveis para apagar definitivamente arquivos de dados de pesquisa constituem um componente crítico para a gestão segura dos dados, que deve estar presente em vários estágios do ciclo de vida dos dados.

Há uma complexidade oculta na eliminação de arquivos, por exemplo, **deletar arquivos armazenados em discos rígidos não previne contra uma possível recuperação desses dados.** Deletar simplesmente remove a referência aos arquivos que, dessa forma,

[51]EYNDEN, Veerle et al. **Managing and data sharing:** best practice for researchers. Colchester: UK Data Archive, 2011. Disponível em: <http://www.data-archive.ac.uk/media/2894/managingsharing.pdf >. Acesso em: 01 out. 2015.

podem ser facilmente restaurados. **Mesmo a reformatação** de discos rígidos não impede a recuperação dos arquivos.

Existem **softwares voltados para a eliminação segura de arquivos** armazenados em discos rígidos que são aderentes aos padrões da área. Peça ajuda à equipe de TI na identificação da melhor ferramenta.

Dispositivos baseados em memória flash, como pen-drives, são construídos de forma diferente dos discos rígidos e as técnicas aplicadas a eles podem não ser confiáveis quando usadas em discos compostos por memória flash. Neste caso, a **destruição física** é o único meio recomendado para apagar os arquivos.

QUAL A FORMA MAIS SEGURA DE ELIMINAÇÃO DE ARQUIVOS DE DADOS?

A forma mais segura de eliminar arquivos é a DESTRUIÇÃO FÍSICA. Trituradores certificados para o nível de segurança desejado devem ser usados para a destruição de papéis e discos de CD/DVD/BLUE RAY.

No final de sua vida útil, discos rígidos de computadores ou discos externos devem ser removidos de seus estojos e eliminados de forma segura por meio de destruição física.

7
PRESERVE SEUS DADOS

Uma etapa crítica para a preservação dos seus **dados de pesquisa é a submissão deles a um arquivo capaz de mantê-los por longo prazo**, tais como um **repositório de dados** ou um **centro de dados**. Esses locais possuem recursos técnicos e gerenciais para fazer a **preservação digital** dos dados e submetê-los a processos mais elaborados conhecidos como **curadoria de dados de pesquisa**.

A preservação não é um processo que se inicia quando a coleção de dados está completa! Na verdade, é um **processo continuo de gestão** que permeia todo o ciclo de vida do projeto de pesquisa e se inicia com o planejamento dos dados que serão obtidos. Além do mais, a preservação ocorre em duas escalas temporais – de curto e de longo prazo – cada uma delas envolve diferentes enfoques e decisões.

- O **armazenamento de curto prazo** é a forma mais comum e também a forma mínima de preservação. Ele se efetiva pela gestão de arquivos de ***backup*** que podem ser criados manualmente ou por sistemas automatizados de armazenamento. Conforme visto no capítulo anterior, os *backups* são cópias do arquivo original, são como uma fotografia do dado num determinado instante de um ciclo em andamento. **Eles são exigidos para restaurar arquivos que tenham sido corrompidos, perdidos, alterados irreversivelmente ou destruídos**.

- O **armazenamento de longo prazo**, por sua vez, tem requisitos diferentes. Um conjunto de dados que está submetido a um processo de arquivamento é geralmente um conjunto de registros que não sofrerá mais mudanças, é como uma fotografia histórica. Ele será **preservado para usos futuros**, portanto ele precisa ser recuperável e estar armazenado em **formatos de arquivos estáveis**, amplamente usados, padronizados e abertos. **É desejável também que o dado esteja acessível livremente**.

O armazenamento de longo prazo depende de **infraestruturas tecnológicas duradouras** que geralmente estão fora do escopo da maioria das organizações de pesquisa, dessa forma, **os repositórios compartilhados se configuram como a melhor solução**.

Identificar um repositório apropriado aos dados do seu projeto é um passo essencial, pois essa decisão pode afetar a coleção e a descrição dos dados. [52]

[52] WIGGINS, Andrea et al. **Data management guide for public participation in scientific research**. Albuquerque, NM: DataONE, 2013. Disponível em: <https://www.dataone.org/sites/all/documents/DataPolicyGuide.pdf>. Acesso em: 10 out. 2015.

7.1
SUBMETA SEUS DADOS A UM REPOSITÓRIO DE DADOS

Você pode **compartilhar os seus dados informalmente** enviando-os por e-mail aos colegas interessados ou postando-os num *website* ou *blog*, porém esses métodos de compartilhamento tornam difícil a tarefa de descobri-los e acessá-los, principalmente por longo prazo.

Armazená-los em mídias portáteis, computador pessoal ou mesmo num servidor da sua instituição expõe seus dados às fragilidades intrínsecas da informação digital – como obsolescência tecnológica, instabilidade das mídias e alterações indevidas.

Os **repositórios de dados de pesquisa são banco de dados** que recebem, gerenciam e disseminam dados e conjuntos de dados de pesquisa. Eles constituem o lugar mais apropriado para que seus dados sejam preservados e possam ser recuperados, acessados e citados por outros pesquisadores, ou seja, tenham visibilidade em escala mundial.

COMO IDENTIFICAR UM REPOSITÓRIO ADEQUADO PARA OS MEUS DADOS?

Os diretórios de repositórios de dados são ferramentas importantes para identificar os repositórios mais apropriados para abrigar os seus dados. Eles armazenam informações sobre os repositórios em operação, permitindo que eles sejam mais facilmente localizados pelas comunidades interessadas, e dessa forma dando mais visibilidade aos resultados da sua pesquisa. Um diretório importante é o **Re3data – REGISTRY OF RESEARCH DATA REPOSITORIES**[53]

Consulte também seus colegas e a sua biblioteca sobre qual é o melhor repositório para a disciplina que você atua e para o dado que sua pesquisa recolheu ou gerou. Verifique também as orientações estabelecidas no PGD (Plano de Gestão de Dados). Considere ainda as exigências sobre depósito dos dados que a sua instituição e/ou financiador da pesquisa tenha, porventura, estabelecido.

Verifique também se **sua instituição possui um repositório digital** que seja apropriado para depositar seus dados. Em muitos casos o depósito em repositórios digitais locais faz parte da política mandatória das instituições de pesquisa.

[53] <http://www.re3data.org/>

QUAIS OS PRINCIPAIS TIPOS DE REPOSITÓRIOS E CENTROS DE DADOS?

Há diferentes tipos de repositórios e centros de dados que podem hospedar seus dados:

- **REPOSITÓRIOS INSTITUCIONAIS**
  Repositórios orientados para materiais produzidos por membros de uma instituição de pesquisa específica. Este tipo de repositório geralmente está aderente a protocolos que permitem a interoperabilidade com outros repositórios. Por exemplo: CARPE DIEN do Instituto de Engenharia Nuclear da CNEN.

- **REPOSITÓRIOS TEMÁTICOS**
  Repositórios dedicados a dados de uma determinada área de conhecimento.

- **REPOSITÓRIOS DE DADOS DE ARTIGOS DE PERIÓDICOS**
  Repositórios vinculados a um ou mais periódicos científicos. Neste tipo de repositório os autores depositam os dados relativos aos artigos publicados pelos periódicos. Por exemplo: **DRYAD** (https://www.datadryad.org) que armazena dados de artigos revisados por pares de mais de 150 periódicos da área de biociências.

- **REPOSITÓRIOS GOVERNAMENTAIS**
  Repositórios de dados apoiados por agências governamentais. Por exemplo: **DATA.GOV** (www.data.gov).

A despeito da variedade de repositórios de dados atualmente existente, são relativamente poucos os que são apropriados para uma dada coleção de dados. Alguns repositórios espelham ou duplicam recursos de dados, agregando conjunto de dados depositados em outros (geralmente menores) repositórios, dessa forma assegurando preservação e capacidade de eles serem recuperados.

QUE OUTROS PONTOS DEVEM SER CONSIDERADOS NA COMPARAÇÃO ENTRE REPOSITÓRIOS?

- **PRIVACIDADE E SENSIBILIDADE**
  Como são tratadas as questões de privacidade e sensibilidade dos dados.

- **CONTROLE DE ACESSO**
  Quais são as opções de controle de acesso relativas às informações sobre o uso dos dados do projeto, de privacidade e de políticas de compartilhamento.

- **POLÍTICAS DE ATRIBUIÇÃO E FISCALIZAÇÃO**

- **USO DOS DADOS**
  Disponibilidade de informações sobe o uso dos dados, com o objetivo de demonstrar o impacto do projeto (por exemplo, número de *downloads*).

- **POLÍTICA DE BACKUP**
  Se o repositório tem uma política explícita de backup.

- **CUSTOS**
  Se existem custos associados ao uso do repositório.

TODOS OS MEUS DADOS DEVEM SER PRESERVADOS?

Após identificar um repositório adequado para armazenamento de longo prazo para as suas coleções de dados, **o próximo passo é selecionar os dados que deverão ser arquivados**. **Nem todos os dados precisam ser preservados e compartilhados**. Dessa forma, identificar os dados de maior valor pode tornar mais simples a documentação e o depósito. Por outro lado, existem muitos usos não previstos para os dados, o que complica muito a decisão do que preservar.

Dependendo dos produtos do seu projeto de pesquisa, arquivar múltiplos conjuntos de dados pode ser a melhor escolha. Normalmente os dados que são depositados em repositórios são dados derivados que passaram por um processamento mínimo (por exemplo: limpeza de dados); mas em alguns casos **dados brutos e/ou dados analisados podem ser mais apropriados para a preservação**. Para projetos em andamento, a melhor opção é seguir as orientações do administrador do repositório no que diz respeito a que conjuntos de dados preservar e com que frequência atualizá-los.

DEVO ORIENTAR O PLANO DE GESTÃO PARA AS EXIGÊNCIAS DO REPOSITÓRIO?

Sim. Moldando o seu plano de gestão na direção de atender as exigências de um repositório específico onde você deseja submeter seus dados, ou sua instituição ou órgão financiador indique ou exija, aumenta a probabilidade dos dados serem aceitos.

VERIFIQUE AS EXIGÊNCIAS DO REPOSITÓRIO EM RELAÇÃO À DESCRIÇÃO DOS DADOS

Verifique as exigências em termos de documentação dos dados, **padrões de metadados, formatos de arquivo e possíveis restrições de uso**, por exemplo, direitos sobre propriedade intelectual.

8
COMPARTILHE SEUS DADOS

Dados de pesquisa são recursos informacionais valiosos que geralmente requerem muito **tempo e dinheiro para serem produzidos**. Se eles forem apropriadamente gerenciados podem ser **usados agora e no futuro por outros pesquisadores** evitando a duplicação de esforços. Além do mais, os dados de pesquisa podem ser **reusados** por outras disciplinas, isto por que muitos deles têm um valor extraordinário que **extrapola o seu propósito original**, podendo ser interpretados em contextos diferentes do que aqueles para os quais foram gerados ou coletados incialmente.

Em um estudo de 2010 sobre dados abertos no Reino Unido, os pesquisadores identificaram os seguintes benefícios para eles mesmos:

- Aumento na eficiência da pesquisa, por exemplo, evitando a duplicação de esforços através da pronta disponibilidade de ferramentas de pesquisa, protocolos e exemplos de boas práticas, pela redução do custo de formação de coleção de dados e pela promoção e adoção de padrões abertos;
- Incentiva a investigação e o debate científico;
- Promove a inovação e potencializa novos usos para os dados;
- Possibilita novas formas de colaboração entre usuários e criadores de dados;
- Maximiza a transparência e a prestação de contas;
- Permite uma avaliação mais precisa das descobertas científicas;
- Colabora para o aperfeiçoamento e a validação dos métodos científicos;
- Evita o custo da duplicação de coleções de dados;
- Aumenta a visibilidade do impacto e a visibilidade dos resultados de pesquisa;
- Contribui para dar os devidos créditos ao criador dos dados;
- Oferece um recurso importante para a educação e a capacitação.

O QUE É NECESSÁRIO PARA COMPARTILHAR MEUS DADOS DE PESQUISA?

Para que dados de pesquisa possam ser efetivamente compartilhados, é necessário uma série de cuidados que se estendem por cada estágio do processo de desenvolvimento dos dados, incluindo:

- **DESCRIÇÃO**
  É necessário que os dados estejam **descritos e documentados em termos de processos, conteúdo e caraterísticas.** Isso é realizado fundamentalmente por **metadados** e outros processos de descrição, como **caderno de laboratório**.

- **ARQUIVAMENTO E ARMAZENAMENTO**
  Os dados precisam estar **depositados em um local confiável onde eles possam ser localizados, acessados, compartilhados e reusados**.

- **PRESERVAÇÃO**
  Os dados devem estar em **formatos de arquivo estáveis**, padronizados e abertos, e registrados em mídias duradouras que facilitem o reuso a longo prazo; devem também ser submetidos a processos que permitam contornar os problemas de obsolescência tecnológica e fragilidade das mídias digitais, como, por exemplo, processos de migração.

- **DESCOBERTA**
  Os dados devem ser passíveis de serem descobertos, para tal é necessário que as **informações sobre as coleções de dados disponíveis sejam amplamente disseminadas** por meio, por exemplo, de publicações acadêmicas, *data clearinghouse* e portais de agregação de dados.

ONDE POSSO ARQUIVAR MEUS DADOS PARA QUE ELES SEJAM COMPARTILHADOS POR OUTROS PESQUISADORES?

Existem várias maneiras de você disponibilizar os seus dados para compartilhá-los com ouros pesquisadores:

- **DEPOSITAR EM REPOSITÓRIO DE DADOS OU CENTRO DE DADOS ESPECIALIZADO**
  Depositando-os em um **repositório de dados**, **centro de dados especializado** ou banco de dados; estes locais podem fornecer um ambiente seguro para os dados.

- **SUBMETER A UM PERIÓDICO CIENTÍFICO**
  Submetendo-os a um **periódico científico,** para complementar o conteúdo de uma publicação acadêmica (muitos periódicos disponibilizam arquivos para depósito de dados referentes aos artigos publicados)

- **DEPOSITAR EM REPOSITÓRIO INSTITUCIONAL OU SISTEMA DE AUTOSUBIMISSÃO**
  Depositando-os em um **repositório institucional** da sua instituição;

- **DISSEMINAR VIA WEBSITE**
  Tornando-os disponíveis online **via *website*** de um projeto ou da sua instituição;

- **DIVULGAR INFORMALMENTE ENTRE OS PARES**
  Tornando os dados **disponíveis** entre colegas pesquisadores através de distribuição informal.

QUAL É A MELHOR OPÇÃO DE ARQUIVAMENTO VOLTADA PARA O COMPARTILHAMENTO?

A escolha vai depender da sua área de pesquisa e do ambiente em que ela se realiza. As opções que se apresentam têm vantagens e desvantagens, porém a opção mais consistente permanece sendo **os centros e repositórios de dados.**

A disponibilização via ***website*** oferece uma forma de fácil e imediata de armazenamento e disseminação, mas pode ser pouco sustentável e apresentar dificuldade de controlar quem usa os dados e como; além do mais, não pressupõe a existência de mecanismos de preservação de longo prazo.

Os **repositórios institucionais**, por sua vez, podem não ser capazes de gerenciar os dados de pesquisa mais complexos, pois estão voltados para o arquivamento de materiais convencionais; as políticas de acesso e preservação e as possibilidades de outros pesquisadores descobrirem os dados têm que ser avaliadas.

A **divulgação informal** entre os pares tem um alcance restrito e depende do ciclo de contatos dos autores; além disso, torna a gestão do acesso uma tarefa complicada e não assegura a disponibilidade dos dados por longo prazo.

Os **centros de dados e os repositórios de dados** são os locais mais apropriados para o arquivamento voltado para o compartilhamento e para a preservação. Entretanto, nem sempre os repositórios e centros de dados aceitam todos os dados que são submetidos para arquivamento, pois, como todo arquivo tradicional, geralmente aplica critérios para **avaliar e selecionar os dados que serão aceitos para preservação**.

QUAIS SÃO AS VANTAGENS DE DEPOSITAR MEUS DADOS EM CENTROS E REPOSITÓRIOS DE DADOS?

As vantagens de depositar os seus dados nesses arquivos são as seguintes:

- **QUALIDADE**
  Assegura que os dados estejam em conformidade com os **padrões de qualidade**;

- **PRESERVAÇÃO DIGITAL**
  Garante a **preservação digital** de longo prazo dos dados;

- **SEGURANÇA**
  Mantém os dados em **ambiente seguro**;

- **BACKUP**
  Providencia **backups** regulares dos dados;

- **DESCOBERTA**
  Viabiliza a **descoberta *online*** dos dados via catálogo de dados;

- **FORMATOS CONHECIDOS**
  Acesso aos dados em formatos populares;

- **DIREITOS**
  Mantém acordo de licenciamento que reconhece os **direitos sobre os dados**;

- **CITAÇÃO**
  Disponibiliza mecanismos de citação que reconhece a **autoria** dos dados;

- **REUSO**
  Promove o **uso** e **reuso** de dados para muitos usuários;

- **GESTÃO DE ACESSO**
  Gerencia o **acesso aos dados** e as **consultas dos usuários** em nome do proprietário dos dados.

QUEM SE BENEFICIA COM O COMPARTILHAMENTO DOS MEUS DADOS?

Qual a razão de se dispender um esforço extra para que seja possível compartilhar dados de pesquisa? São muitas as razões e o benefício atinge a vários segmentos da sociedade.

- **AGÊNCIA DE FOMENTO À PESQUISA**
  As organizações que financiam as pesquisas científicas têm como **obrigação maximizar os seus investimentos**, nessa direção, o compartilhamento de dados amplia o valor dos investimentos na medida em que diferentes pesquisadores podem reusar os dados produzidos/coletados originalmente por outros projetos, evitando a duplicação de esforços e recursos.

- **COMUNIDADE CIENTÍFICA**
  A disponibilidade dos dados permite que os **revisores verifiquem com mais precisão a qualidade e a autenticidade dos produtos de um projeto de pesquisa**, como por exemplo, os artigos de periódicos. Além do mais, o acesso às pesquisas relacionadas permite que os membros da comunidade científica reproduzam, comparem e avaliem métodos e resultados de uma forma precisa.

- **PESQUISADOR**
  Quando um pesquisador compartilha seus dados **ele ganha reconhecimento como autor e como uma fonte confiável de conhecimento na a área em que atua**. Ele pode, dessa forma, ser citado e referenciado e obter os créditos acadêmicos correspondentes. Quando os dados são expostos, o *feedback* da comunidade pode ser usado para melhorar a qualidade e a apresentação dos dados.

- **CIDADÃO COMUM**
  Tem acesso aos produtos do trabalho de pesquisa financiados com verba pública de forma transparente.

QUANDO OS DADOS DEVEM SER COMPARTILHADOS?

Os padrões seguidos pelas comunidades científicas e as **políticas mandatórias** estabelecidas pelas agências financiadoras de pesquisa variam de acordo com as disciplinas e com os tipos de dados. A maioria das agências de fomento exige que os dados sejam disponibilizados "**dentro de um tempo razoável**"; algumas agências determinam um período de tempo especifico para que os dados sejam compartilhados – por exemplo, 2 ou 3 anos após os dados ter sido coletados ou até que os resultados baseados nos *data sets* sejam aceitos para publicação.

Período de embargo

Além disso, muitas agências também permitem **períodos de embargo** (período de tempo no qual os dados não são disseminados) por razões políticas, comerciais ou por **processos de patentes**.

QUE ASPECTOS ÉTICOS E POLÍTICOS DEVEM SER CONSIDERADOS QUANDO SE COMPARTILHA DADOS?

Quando um pesquisador compartilha dados ou usa dados de outras fontes, ele deve estar ciente das considerações **legais e políticas que afetam o uso e reuso** desses dados.

Quando você disponibiliza seus dados é importante elaborar uma **declaração de direitos de uso** apropriada para dados, que esteja de acordo com a política de sua instituição ou da agência financiadora. Esta declaração deve estar incluída na documentação dos seus dados. Dessa forma, os usuários estarão cientes das **condições de uso** desses dados. A declaração de direitos de uso deve incluir quais são **os usos apropriados dos dados**, como contatar o autor dos dados e ainda como identificar a fonte desses dados.

Existem três áreas principais que necessitam ser endereçadas quando se produz dados que podem ser compartilhados:

- **PRIVACIDADE E CONFIDENCIALIDADE**
  Os dados devem estar aderentes às políticas de privacidade e confidencialidade de sua instituição;

- **COPYRIGHT E PROPRIEDADE INTELECTUAL**
  Dados não podem ser submetidos às leis de *copyright*. Se você usa dados de outras fontes assegure-se de que você tem a permissão apropriada, especialmente para dados que têm múltiplos proprietários ou *copyright* layers. Atente para o fato de que a documentação sobre o contexto da coleção de dados pode estar protegida por *copyright*.

- **LICENCIAMENTO**

Dados de pesquisa podem ser licenciados, portanto a forma como você licencia seus dados pode determinar suas possibilidades de uso por parte de outro pesquisador. Por exemplo, o **Creative Zero License** proporciona um acesso bastante amplo.

9

FORMATE SEUS DADOS[54]

Dados de pesquisa se apresentam numa grande variedade de formatos: textual, numéricos, multimídia, imagens, simulações, modelos, linguagem de software, formatos específicos de disciplinas e de instrumentos etc. Isto torna mais complexa a preservação dos dados.

Os formatos em que os dados de pesquisa são criados geralmente dependem de como os pesquisadores planejam analisar os dados, o *hardware* usado, a disponibilidade de *software*, ou podem ainda ser determinados por padrões específicos de uma disciplina. Porém, para assegurar a usabilidade dos dados por longo prazo é necessário que se considere quais são os formatos de arquivos e *software* mais apropriados.

O uso de formatos de arquivos padronizados e abertos assegura que os dados possam ser usados e reusados pelo tempo que for necessário. Dessa forma, se torna importante criar ou converter os formatos de arquivos dos dados para um elenco pré-determinado pela instituição que possa ser mais facilmente gerenciado.

FORMATOS DE ARQUIVO

Toda a informação digital é planejada para ser interpretada por um programa de computador. Sem esse programa a informação não pode ser compreendida e inexiste. Isto significa que os dados digitais são ameaçados pela obsolescência tecnológica do ambiente de hardware e de software necessários à interpretação deles.

Mesmo considerando a compatibilidade retrospectiva de muitos pacotes de software - que permite que dados criados em versões anteriores sejam lidos em versões atuais do software - e a interoperabilidade entre softwares concorrentes, a opção mais segura para garantir o acesso de longo prazo é converter os dados para formatos padronizados. Dessa forma, os dados podem ser interpretados por vários programas, e ficam mais apropriados para o intercâmbio e preservação.

Isto significa usar formatos abertos e padronizados como o OpenDocument Format (ODF), ASCII, XML, valores separados por vírgula, formatos delimitados por tab. Alguns formatos proprietários tais como o MS Rich Text Format, MS Excel, SPSS são largamente utilizados e provavelmente serão acessíveis por um tempo razoável, mas não ilimitado.

Os pesquisadores podem usar os *software* e formatos de dados mais apropriados às analises que foram planejadas; uma vez que essas análises foram completadas e os dados estão sendo preparados para arquivamento, o pesquisador deve considerar

[54] UK DATA ARCHIVE. **Craate & Manage Data**: formatting your data. Disponível em: <http://www.data-archive.ac.uk/create-manage/format>. Acesso em: 01 out. 2015.

converter os dados para formatos padronizados, intercambiáveis e estáveis por longo prazo, de forma a preservar o potencial de uso dos dados para o futuro.

QUE TIPOS DE FORMATOS DE ARQUIVO DEVO USAR PARA GARANTIR A LONGEVIDADE DOS MEUS DADOS?

Os formatos de arquivo que são mais prováveis de poder ser acessados no futuro possuem as seguintes características:

- **NÃO PROPRIETÁRIOS**;
- **ABERTOS E PADRONIZADOS** (têm a documentação disponível livremente);
- **USADOS COMUMENTE PELA COMUNIDADE DE PESQUISA**;
- **USA CARACTERES DE CODIFICAÇÃO PADRONIZADOS** (ASCII, UTF-8);
- **SEM COMPRESSÃO**.

NEM SEMPRE MEUS DADOS PODEM ESTAR EM ARQUIVOS ABERTOS E PADRONIZADOS, O QUE FAZER PARA TORNÁ-LOS DURADOUROS?

Os pesquisadores devem **usar os dados nos formatos mais convenientes** e os softwares de acordo com as análises que foram planejadas. Uma vez que a análise dos dados foi completada e os dados estão preparados para o arquivamento, o pesquisador deve considerar **convertê-los para um formato padronizado, intercambiável e mais duradouro**.

A CONVERSÃO PARA FORMATOS PADRONIZADOS PODE CAUSAR PERDAS?

Sim! Quando os dados são convertidos de um formato de arquivo para outro, seja por meio de exportação ou por meio de *software* de conversão de dados, algumas alterações podem ocorrer com os dados. Portanto, depois da conversão eles devem ser checados para detectar possíveis erros ou mudanças causadas pelo processo de exportação. Por exemplo:

- **PARA DADOS TEXTUAIS**, características como *highlighting*, negrito, notas de rodapé podem ser perdidas;
- **PARA DADOS ESTATÍSTICOS**, planilhas ou base de dados, alguns dados ou metadados internos tais como definição de dados ausentes, números decimais, fórmulas, podem ser perdidos ou dados podem ficar truncados na conversão.

QUAIS SÃO OS FORMATOS DE ARQUIVO RECOMENDADOS PARA COMPARTILHAMENTO, REUSO E PRESERVAÇÃO DE LONGO PRAZO?

A tabela abaixo apresenta os arquivos geralmente aceitos para depósitos em arquivos de dados.

TIPOS DE DADOS	ARQUIVOS DE DADOS RECOMENDADOS	OUTROS FORMATOS ACEITÁVEIS
DADOS QUANTITATIVOS TABULAR COM METADADOS EXTENSIVOS	• ***SPSS portable format* (.por)** • ***Delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information*** • ***Some structured text or mark-up file containing metadata information*, por exemplo, DDI XML file**	• **Formatos proprietários de pacotes estatísticos, ex. SPSS (.sav), Stata (.dta)** • **MS Access (.mdb/.accdb)**
DADOS QUANTITATIVOS TABULAR COM METADADOS MÍNIMOS	• **CSV - Valores separados por vígula (.csv)** • ***Tab-delimited file* (.tab)** **Incluindo texto delimitado por um dado conjunto de caracteres com declaração de definição de dados SQL quando apropriado.**	• **Texto delimitado por um dado conjunto de caracteres – somente caracteres que não estão presentes nos dados devem ser usados como delimitador (.txt)** • **Formatos populares, ex. MS Excel (.xls/xlsx), MS Access (.mdb/.accdb), dBase (.dbf) e *Open Document Spreadsheet* (.ods)**
DADOS GEOESPACIAIS Dados vetoriais ou raster	• **ESRI Shape file** (essencial: .shp; .shx; dbf/ opcional: .prj; .sbx; .sbn) • **TIFF geo-referenciado** (.tif; .tfw) • **CAD** (.dwg) • **Tabular GIS attribute data**	
DADOS QUALITATIVOS Textual	• **Texto XML** (.xml) **de acordo com DTD** (.txt) **ou esquema** (.xml) • **Rich Text Format** (.rtf) • **Textual plano, UFT-8** (Unicode; .txt)	• **Texto plano ASCII (.txt)** • **HTML (.html; .htm)** • **Formatos proprietaries populares, como MSWord (.doc; .docx)** • **LaTex (.tex)**
IMAGEM DIGITAL	• **TIFF versão 6 sem compressão** (.tif)	• **JPEG (.jpg; .jpeg)** • **TIFF outras versões (.tif; .tifff)** • **JPEG 2000 (.jp2)** • **PDF/A, PDF (.pdf)** • **RAW image format (.raw)**

		• **Arquivos Photoshop (.psd)**
ÁUDIO DIGITAL	• **FLAC - Free Lossless Audio CODEC** (.flac) • **WAVE - Wafeform Audio Format** (.wav) • **MP3 – MPEG-1 Audio Layer 3** (.mp3) (somente para discurso, entrevista, etc)	• **AIFF- Audio Interchange File Format (.aif)** • **MP3 – MPEG-1 Audio Layer 3 (.mp3)**
VÍDEO DIGITAL	• **MPEG-4 High Profile** (.mp4) • **Motion JPEG 2000** (.jp2)	
DOCUMENTAÇÃO e *SCRIPTS*	• **Rich Format Text (.RTF)** • **Open Document Text (.odt)** • **Rich Document Format (.rtf)** • **PDF/A ou PDF (.pdf)** • **HTML (.htm; .html)**	• **plano (.txt)** • **XML acompanhado de DTD ou esquema XML** • **Formatos proprietários populares, como MS Word (.doc; .docx) ou MS Excel (.xls; .xlsx)**

10
GARANTA A QUALIDADE DE SEUS DADOS![55]

A gestão da **qualidade dos dados** é um conjunto de ações que deve permear todo o ciclo de desenvolvimento do projeto de pesquisa. Estas ações asseguram a qualidade dos dados antes deles serem coletados, entrados ou analisados e monitoram e mantém a qualidade dos dados no decorrer do projeto, aumentando o seu nível de confiabilidade e a sua potencialidade de uso, reuso e compartilhamento.

O QUE É GARANTIA E CONTROLE DE QUALIDADE?

Garantia de Qualidade e **Controle de Qualidade**, no contexto da gestão de dados de pesquisa, são termos usados para descrever os procedimentos relacionados à prevenção e à minimização da introdução de erros em conjuntos de dados e à identificação de dados errôneos. Para assegurar a qualidade dos dados e torná-los potencialmente mais úteis para usos atuais e futuros existem várias abordagens que devem ser aplicadas durante todo o ciclo de desenvolvimento de um projeto de pesquisa. Esse conjunto de abordagens e de metodologias pode ser **aplicado antes, durante e depois da coleta de dados**, respeitando as especificidades de cada um desses momentos: as estratégias usadas antes e durante a coleta de dados são chamadas coletivamente de **Garantia de Qualidade** e as aplicadas posteriormente são chamadas de **Controle de Qualidade**.

- **GARANTIA DE QUALIDADE** – refere-se aos processos usados para assegurar que **os melhores dados possíveis serão coletados**.

- **CONTROLE DE QUALIDADE** – é um conjunto de processos para avaliar a **qualidade dos dados após eles serem coletados**. Envolve "limpeza de dados" e a tomada de decisões sobre questões tais como lidar com dados ausentes e valores estimados. O controle de qualidade é considerado mais difícil e exige mais recursos do que a garantia de qualidade, pois – de uma forma geral - é mais fácil prevenir do que reparar problemas, além de ser muito mais barato ao longo do tempo.

A qualidade dos dados depende de muitos fatores, o que significa que esses procedimentos por si só não são capazes de garantir completamente a usabilidade dos dados. Entretanto, um planejamento bem documentado dos procedimentos de garantia

[55] DATAONE. **Tutorials on data management Lesson 5**: Data Quality Control and Assurance <https://www.dataone.org/sites/all/documents/L05_DataQualityControlAssurance.pptx>. Acesso em: 10 out. 2015.

e controle de qualidade aumenta a probabilidade de que os dados possam ser usados e reusados. Portanto, é crítico documentar, com o maior nível de detalhe possível, os processos relacionados à qualidade dos dados, e registrar também qualquer mudança que esses processos tenham sofrido. Esse procedimento beneficia tanto as **pessoas que gerenciam os dados**, quanto **os pesquisadores que precisam utilizá-los**.

A qualidade dos dados é um parâmetro relativo. Assegurar a qualidade dos dados requer conhecimento sobre **os critérios que os dados devem atender em relação aos objetivos e/ou padrões científicos** que se pretende alcançar: a qualidade dos dados é determinada pelo nível de adequabilidade dos dados aos usos pretendidos para eles.

Durante o planejamento, considere não somente quais os mecanismos de qualidade de dados são apropriados para os diferentes estágios do projeto, mas também os custos associados. Além do mais, é de grande importância determinar claramente os **papeis e as responsabilidades** pela garantia de qualidade dos dados em todos os estágios da pesquisa.

QUAIS SÃO OS TIPOS MAIS COMUNS DE ERROS QUE PODEM OCORRER EM UM CONJUNTO DE DADOS?

Em geral, existem dois tipos de erros que podem ocorrer em uma coleção de dados: comissão e omissão:

- **Erro por comissão** é o resultado de dados incluídos de forma incorreta ou imprecisa na *data set*. Isto pode ser causado, por exemplo, por:
    - **Mau funcionamento de um instrumento** que produz resultados incorretos;
    - **Dados que são digitados incorretamente** durante a entrada de dados.
- **Erros por omissão**, por sua vez, são resultados de **dados ou metadados omitidos**. Situações que resultam em erros por omissão acontecem, por exemplo, quando:
    - Os dados são **documentados de forma inadequada** para uso efetivo;
    - Há **erros humanos durante a coleta ou a entrada de dados**, por exemplo, uma medida é esquecida, ou uma linha da planilha é ignorada durante a entrada de dados;
    - Existem **anomalias no campo que afetam os dados**. Se as anomalias que são conhecidas por afetar os dados não são documentadas e reportadas nos metadados, então dados errados podem ser registrados e usados.
    - O aparelho de GPS fica sem bateria e não registra as coordenadas espaciais.

QUAIS AS PRÁTICAS QUE DEVEM SER ADOTADAS DURANTE A COLETA DE DADOS?

Durante a coleta de dados, você deve assegurar que os dados reflitam com fidedignidade os fatos, as respostas, as observações e os eventos que estão sendo registrados.

A qualidade da metodologia usada na coleta dos dados influencia de forma significativa na qualidade dos dados; além do mais, uma documentação detalhada sobre como os dados são coletados fornece as evidências sobre o nível de qualidade desses dados.

Medidas de controle de qualidade durante a coleta de dados podem incluir:

- **Calibrar os instrumentos** checando a precisão, o viés (bias), e/ou a escala de medida;
- **Fazer múltiplas medidas, observações ou coleta de amostras**;
- **Checar a veracidade do registro com um especialista**;
- **Usar métodos e protocolos padronizados para a captura.**

Práticas de controle de qualidade são específicas para cada tipo de dados que está sendo coletado, mas algumas regras gerais podem ser adotadas:

- **Dados coletados por instrumentos**
  - Valores registrados por instrumentos devem ser checados com o objetivo de garantir que esses **valores estão dentro da faixa de sensibilidade do instrumento** e dos **limites da propriedade que está sendo medida**. Por exemplo: concentrações não podem ser < 0; a velocidade do vento não pode ser maior que o anemômetro pode registrar.
- **Resultados analíticos**
  - **Valores medidos em laboratórios** têm que ser checados para assegurar que eles estão dentro do limite de detecção do método analítico e são válidos para o que está sendo medido. Se os valores estão abaixo do limite de detecção, eles têm que ser apropriadamente codificados e qualificados.
  - Qualquer **dado ancilar** usado para avaliar a qualidade dos dados tem que ser descrito e armazenado. Por exemplo: dados usados para comparar leitura de instrumentos contra padrões conhecidos.
- **Observações** (tais como contagens de pássaros ou cobertura vegetal)
  - **Checagem da faixa de ocorrência** e **comparações com valores históricos máximos** ajudarão a identificar valores anômalos que vão exigir investigação complementar.
  - **Comparações entre mensurações atuais e passadas** ajudarão a identificar eventos altamente improváveis. Por exemplo: é improvável que a circunferência de uma árvore irá decrescer de um ano para outro.

QUAIS AS PRÁTICAS QUE DEVEM SER ADOTADAS DURANTE A ENTRADA DE DADOS?

Quando os dados são digitados, transcritos, entrados em uma planilha ou banco de dados, ou codificados, o uso de procedimentos padronizados e consistentes, acompanhados de instruções claras são práticas que irão assegurar a qualidade e evitar a introdução de erros. Isto pode incluir, por exemplo:

- **Entrada dupla –** os dados digitais são teclados por duas pessoas de forma independente; diferenças na entrada podem ser detectadas por programa de computador.

- **Grave e transcreva –** outra maneira de reduzir erros na entrada de dados é gravar você mesmo a leitura dos dados e depois transcrever a partir da gravação.

- **Programa leitor de texto –** você pode usar também um programa que faça a leitura dos dados enquanto você faz a digitação deles no computador.

Se você pretende usar *software* de **planilha** (por exemplo, o MS Excel) ou de **banco de dados** (por exemplo, o MS Access) para registrar os seus dados, você deve projetar antecipadamente uma estrutura voltada para organizar os seus dados e arquivos de dados. Atente para os seguintes pontos:

- Use **terminologia consistente** dentro da base de dados, utilizando, por exemplo, **vocabulários controlados**, **listas de códigos**, **listas de opções,** além do mais, isto vai minimizar o esforço da entrada manual.

- Use **lista de códigos** – entradas codificadas podem ser checadas contra uma lista de valores permitidos, validando os valores que estão sendo entrados.

- **Use regras de validação dos dados –** os *softwares* de banco de dados permitem que você defina regras para validar os valores que podem ser entrados em cada campo, por exemplo, configure o campo para só aceitar dados textuais ou valores numéricos; especifique o número máximo de caracteres ou a faixa de valores que o campo pode aceitar, ou configure o campo para aceitar somente valores únicos.

- **Atomize os dados** – para o caso de você usar planilha, registre somente **um item de informação em cada célula**, pois o lançamento de múltiplos itens de informação numa única célula acarretará problemas durante a fase de análise dos dados; os **valores têm que ser consistentes com o tipo de dado** (inteiro, textual, data/tempo) da coluna na qual eles estão sendo incluídos. Por exemplo: 12-20-2000A não pode ser entrado em uma coluna de datas.

Documente todas as alterações realizadas sobre os dados. Documentar mudanças nos dados pode ser tão simples como criar um **arquivo texto para acompanhar o conjunto de dados**, ou pode envolver o uso de um **programa baseado em script** para correção de erros, de forma que cada passo tomado possa ser claramente documentado.

- Evite **duplicação de esforços na checagem de erros**;

- Se forem cometidos equívocos na edição ou limpeza de dados, bons registros permitem que esses **equívocos possam ser desfeitos**.

QUAIS AS PRÁTICAS QUE DEVEM SER ADOTADAS DEPOIS DA ENTRADA DE DADOS?

Uma vez que os dados foram digitados, medidas básicas de controle de qualidade podem ser tomadas.

- Se os dados foram registrados em planilha ou banco de dados, assegure-se de que eles **estão alinhados na coluna correta**;

- **Verifique valores ausentes, impossíveis e anômalos**. Uma forma de checar esses problemas é ordenar o campo de dados e verificar as discrepâncias.

- Procure por valores discrepantes (*outliers*) – outliers são valores extremos para uma variável que estão fora do modelo estatístico usado para descrever os dados (fora da curva). O objetivo não é eliminar esses valores, mas identificar a possível contaminação dos dados. Esses dados podem ser marcados para posterior investigação.

COMO AMPLIAR O VALOR DOS MEUS DADOS?

Os pesquisadores podem ampliar significativamente o valor dos seus conjuntos de dados para reuso em outros projetos e contextos incluindo variáveis adicionais ou parâmetros que aumentam as possibilidades de aplicações dos dados. Por exemplo: georreferenciar os dados permite que outros pesquisadores apliquem os dados em sistemas de informação geográficos.

11

ÉTICA E CONSENTIMENTO[56]

Coletar, usar e compartilhar dados no âmbito de pesquisas que envolvam pessoas exige que obrigações éticas e legais sejam respeitadas.

Quando a pesquisa envolve obter **dados de pessoas**, o que se espera do pesquisador é que ele mantenha um comportamento pautado por um **rigoroso código de ética**, que seja condizente com os padrões e protocolos recomendados pelas entidades profissionais, instituições de pesquisa e organizações financiadoras de pesquisa e, sobretudo, com a legislação do país concernente a esse aspecto. Este comportamento deve **permear todo o ciclo de pesquisa**, incluindo especialmente a fase de **compartilhamento dos dados**.

Nesse contexto a compreensão de três tipos de dados se torna essencial:

- **DADOS PESSOAIS**
  São dados relacionados a indivíduos vivos, que podem ser identificados a partir desses dados ou a partir desses dados combinados com outras informações.

- **DADOS CONFIDENCIAIS**
  São dados que não estão em domínio público tais como informações sobre negócios, lucros, saúde, detalhes médicos e opiniões políticas, entregues em confiança ou que duas partes concordam em mantê-los confidenciais, isto é, secretos.

- **DADOS PESSOAIS SENSÍVEIS**
  São dados sobre raça, origem étnica, opinião política, religião ou crenças similares, filiação sindical, doença física ou mental, vida sexual, etc.

[56] EYNDEN, Veerle et al. **Managing and data sharing:** best practice for researchers. Colchester: UK Data Archive, 2011. Disponível em: <http://www.data-archive.ac.uk/media/2894/managingsharing.pdf >. Acesso em: 01 out. 2015.

QUAIS OS PRINCÍPIOS CHAVE DA ÉTICA NA PESQUISA QUE TÊM QUE SER CONSIDERADOS NO COMPARTILHAMENTO E ARQUIVAMENTO DE DADOS CONFIDENCIAIS?

- Dever de **confidencialidade** para com os informantes e participantes.
- Dever de proteger os participantes de ofensas, **não divulgando informações sensíveis.**
- Dever de **tratar os participantes como seres inteligentes**, capazes de tomar suas próprias decisões sobre como as informações que eles fornecem podem ser usadas, compartilhadas e tornadas públicas (por meio de consentimento informado).
- Dever de informar aos participantes, antes de obter o consentimento, **como a informação e os dados obtidos serão usados, processados, compartilhados e eliminados.**

OS DADOS DE PESQUISA CONFIDENCIAIS E SENSÍVEIS PODEM SER COMPARTILHADOS?

Mesmo os dados de pesquisa sensíveis e confidenciais podem ser compartilhados ética e legalmente. Isso requer que os pesquisadores atentem, desde o começo da pesquisa, para três aspectos de muita importância.

- Na obtenção do **consentimento informado**, inclua previsão para o compartilhamento de dados.
- Quando necessário, **proteja a identidade** das pessoas via **anonimização** dos dados.
- Use mecanismos de **controle de acesso** aos dados.

QUAIS SÃO AS EXIGÊNCIAS DOS REPOSITÓRIOS EM RELAÇÃO ÀS QUESTÕES DE CONFIDENCIALIDADE E DE DIVULGAÇÃO DOS DADOS?

Os repositórios e centros de dados geralmente exigem que os depositantes de dados de pesquisa assegurem que os dados cumpram as exigências de confidencialidade e de não divulgação dos dados coletados a partir de assuntos que envolvam seres humanos. **Em muitos casos os repositórios podem alterar os dados sensíveis para criar dados anonimizados.**

O QUE É PERÍODO DE EMBARGO?

Período de tempo no qual o acesso e o uso dos dados para certos tipos de usuários podem estar restritos com o objetivo de proteger pesquisadores, pessoas e organizações. Alguns repositórios têm a capacidade técnica de postergar o acesso aos dados até que o conteúdo tenha sido aprovado para divulgação pública. Sobre esse período é necessário considerar:

- Os acordos sobre o período de embargo dos dados precisam ser **estabelecidos de comum acordo entre o repositório e seus depositantes**.
- Alguns repositórios podem tornar **acessíveis os metadados sobre os dados** que estão embargados ou que tenham o acesso restrito por algum motivo.
- Alguns repositórios apresentam a possibilidade de **liberar automaticamente** os dados assim que o embargo termina.

AS REGRAS SOBRE PROTEÇÃO DE DADOS SE APLICAM A TODOS OS DADOS?

Dados coletados a partir de/ou sobre pessoas podem conter informações sensíveis ou confidenciais. Porém, isso não significa que todos os dados obtidos pela pesquisa são pessoais e confidenciais. A legislação sobre proteção de dados se aplica unicamente a dados pessoais ou dados pessoais sensíveis e não sobre toda a coleção de dados de pesquisa ou a dados anonimizados.

11.1
TERMO DE CONSENTIMENTO INFORMADO

Consentimento informado se refere ao processo de comunicação que permite que um indivíduo faça escolhas informadas sobre sua participação em uma pesquisa. Um acordo de consentimento informado fornece as informações necessárias sobre a pesquisa e serve como um compromisso formal para que uma pessoa participe voluntariamente de uma proposta de pesquisa. Uma descrição de como a confidencialidade do participante será protegida deve estar incluída no acordo de consentimento informado[57].

Dados pessoais obtidos a partir de informações de pesquisa nunca devem ser divulgados, a menos que um respondente conceda um consentimento específico para isso, preferencialmente por escrito.

[57] ICPSR. **Guide to Social Science Data Preparation and Archiving**. 2012. Disponível em: <http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf>. Acesso em: 10 out. 2015.

11.2
ANONIMIZAÇÃO DOS DADOS

Os dados obtidos de pesquisas que envolvam pessoas podem ser compartilhados com outros pesquisadores ou arquivados em repositórios. Para isso, os dados precisam passar por processos que impossibilitem que indivíduos, e também organizações e negócios, possam ser identificados.

Nessa direção, a **anonimização** compreende uma série de procedimentos a que devem ser submetidos os dados que contém informações pessoais; isso evita que eles revelem a identidade de indivíduos e impede também que esses dados estejam vinculados a outros e diferentes dados que possam revelar indiretamente a identidade dos indivíduos envolvidos. **Anonimizar** dados de pesquisa pode tomar tempo e, portanto, **custar caro**. Planejar com antecedência pode ajudar a reduzir custos.

QUAIS SÃO AS FORMAS DE IDENTIFICAR UMA PESSOA?

- **IDENTIFICAÇÃO DIRETA**
  São dados tais como NOMES, ENDEREÇOS, CÓDIGO POSTAL, TELEFONE OU FOTOS;

- **IDENTIFICAÇÃO INDIRETA**
  São informações que quando vinculadas a outras fontes de informação disponíveis publicamente podem identificar uma pessoa, por exemplo, informações sobre LOCAL DE TRABALHO, OCUPAÇÃO ou valores excepcionais de características como SALÁRIO ou IDADE.

A identificação direta é usualmente coletada como parte do processo de administração da pesquisa e geralmente não é uma informação essencial para a pesquisa e pode, portanto, ser facilmente removida da coleção de dados.

AS OBRIGAÇÕES ÉTICAS DE QUEM REUSA DADOS SÃO IDÊNTICAS ÀS DO PESQUISADOR PRIMÁRIO?

SIM. Da mesma forma como os usuários primários, os pesquisadores que reusam dados de pesquisa têm as mesmas obrigações legais e éticas de não divulgar informações confidenciais.

12
COPYRIGHT

Copyright é um direito de propriedade intelectual atribuído automaticamente ao criador. Sua aplicação impede que cópias e publicações de uma obra original sejam realizadas sem autorização prévia do detentor dos direitos sobre a obra. A maioria dos produtos de pesquisa – incluindo planilhas, publicações, relatórios e programas de computador – se enquadra como obras literárias e é, portanto, protegida pelas leis de *copyright*. Entretanto, fatos não podem ser protegidos por *copyright*. Mas é importante assinalar que ***copyright* se aplica também aos dados de pesquisa**, e é um **item importante a ser considerado na criação, compartilhamento e reuso de dados**.

Quando dados de pesquisa são compartilhados ou arquivados, o detentor original do *copyright* retém os diretos sobre esses dados. Um repositório ou centro de dados não pode arquivar dados a menos que os detentores dos direitos sobre esses dados sejam identificados e concedam explícita permissão para que os dados sejam compartilhados. Por sua vez, os usuários secundários precisam obter liberação do *copyright* antes que os dados sejam reproduzidos.

A QUEM PERTENCEM OS DIREITOS SOBRE OS DADOS?

Os pesquisadores que criam os dados geralmente detêm os direitos associados a esses dados. Nessa direção, o criador é automaticamente o primeiro proprietário do *copyright* dos dados que ele coletou, a menos que exista um contrato que atribua o *copyright* de forma diferente, ou exista um documento que transfira formalmente os direitos atribuídos ao criador.

No ambiente de uma instituição acadêmica, em tese, o empregador é o proprietário do *copyright* de uma obra realizada durante o período em que o pesquisador está trabalhando na instituição. Entretanto, muitas instituições atribuem o *copyright* dos materiais de pesquisa, dados e publicações – ou seja, resultados de pesquisa – aos pesquisadores que os criaram. **Dessa forma, você deve verificar qual a política de atribuição de *copyright* adotada por sua instituição.**

A QUE PODE SER ATRIBUÍDO *COPYRIGHT*?

Para o *copyright* ser aplicado, a obra tem que ser original e fixada em um suporte material, por exemplo, estar escrita ou gravada. Não existe *copyright* de ideias ou de discursos não gravados; fatos também não podem ser protegidos por *copyright*. Se um pesquisador coleta dados por meio de entrevista e grava ou transcreve a fala do

entrevistado, o pesquisador detém o *copyright* desses registros. Além do mais, cada entrevistado é um autor de suas palavras na entrevista.

O QUE FAZER NO CASO DE PESQUISAS COLABORATIVAS?

No caso de pesquisas colaborativas ou de dados derivados, **o *copyright* pode ser atribuído em conjunto para vários pesquisadores ou instituições**. Nessa direção, você deve estar atento à correta atribuição de *copyright* às coleções de dados que foram criadas a partir de uma variedade de fontes.

Você deve considerar também os direitos associados aos **materiais originais** no caso de você utilizar **representações digitais** de textos, imagens, ou gravações analógicas.

O REUSO DE DADOS ESTÁ SOB *COPYRIGHT*?

Sim. Usuários que reusam os dados – usuários secundários – têm que obter liberação do *copyright* do detentor dos direitos sobre os dados, antes dos dados serem reproduzidos.

Dados compartilhados por meio de centros de dados

Quando os dados são compartilhados por meio de um *data center*, o pesquisador ou o criador dos dados mantém os direitos sobre os dados. O centro está licenciado por esses autores para processar e prover acesso aos dados.

Efetivamente, o centro não tem direitos sobre os dados, a menos que todos os detentores de direitos sejam identificados e deem permissão para que os dados sejam arquivados e compartilhados. Os centros de dados geralmente especificam como os dados devem ser reconhecidos e citados, seja dentro dos registros de metadados do conjunto de dados, ou num documento de licença de uso dos dados.

Dados submetidos a um periódico científico

Quando dados de pesquisa são submetidos a um periódico científico para suplementar uma publicação – por exemplo, um artigo -, o pesquisador precisa verificar se o editor espera que o *copyright* seja transferido.

O CONCEITO DE "USO JUSTO" (*FAIR USE*) SE APLICA AOS DADOS DE PESQUISA?

Sim. Os dados podem ser copiados para ensino não comercial ou para propósito de pesquisa sem infringir o *copyright*, providenciado que o proprietário dos dados seja declarado. Uma declaração deve dar crédito à fonte dos dados usados, ao distribuidor dos dados e ao detentor do *copyright*.

QUE TIPOS DE LICENÇA POSSO USAR PARA OS MEUS DADOS?

Alguns pesquisadores licenciam seus ativos intelectuais usando o conceito da licença **Creative Commons**[58], que permite que o pesquisador comunique os direitos que ele deseja manter e os que ele pode renunciar quando outros pesquisadores reusam esses ativos. Porém, a licença Creative Commons não é adequada para dados. Outras licenças com objetivos similares são mais apropriadas. Um exemplo é a licença **Open Data Commons**[59].

[58] <http://creativecommons.org/>
[59] <http://opendatacommons.org/>

REFERÊNCIAS BIBLIOGRÁFICAS

BALL, Ales. **A review of data management lifecycle models**. Batth, UK: University of Bath, 2012. Disponível em: <http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf>. Acesso em: 01 out. 2015.

BORGMAN, Cristine. Research data: who will share what, with whom, when, and why? In: CHINA--NORTH AMERICAN LIBRARY CONFERENCE, 5., 2010, Beijing. Disponível em: <http://works.bepress.com/borgman/238/>. Acesso em: 10 out. 2015.

DATAONE. **Tutorials on data management Lesson 03**: Data Management Planning. Disponível em: <https://www.dataone.org/sites/all/documents/L03_DataManagementPlanning.pptx>. Acesso em: 10 out. 2015.

_____. **Tutorials on data management Lesson 05**: Data Quality Control and Assurance. Disponível em: <https://www.dataone.org/sites/all/documents/L05_DataQualityControlAssurance.pptx>. Acesso em: 10 out. 2015.

_____. **Tutorials on data management Lesson 06:** Data Protection and Backups. Disponível em: <https://www.dataone.org/sites/all/documents/L06_DataProtectionBackups.pptx >. Acesso em: 01 out. 2015.

_____. **Tutorials on data management Lesson 7:** Metadata. Disponível em: <https://www.dataone.org/sites/all/documents/L07_Metadata.pptx>. Acesso em: 01 out. 2015.

DCC. **Disciplinary Metadata**. Disponível em: <http://www.dcc.ac.uk/resources/metadata-standards>. Acesso em: 10 out. 2015.

DCC. **General Research Data**. Disponível em: <http://www.dcc.ac.uk/resources/subject-areas/general-research-data>. Acesso em: 10 out. 2015.

EYNDEN, Veerle et al. **Managing and data sharing:** best practice for researchers. Colchester: UK Data Archive, 2011. Disponível em: <http://www.data-archive.ac.uk/media/2894/managingsharing.pdf>. Acesso em: 01 out. 2015.

EUROPEAN COMISSION. **Guidelines on data management in horizon 2020**. Dec. 2013. Disponível em: <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf>. Acesso em: 01 out. 2015.

GREEN, Ann; MACDONALD, Stuart; RICE, Robin. **Policy-making for research data in Repositories**: a guide. May 2009. Disponível em: <https://www.coar-repositories.org/files/guide.pdf>. Acesso em: 01 out. 2015.

HOOK, Les A. et al. **Best Practices for Preparing Environmental Data Sets to Share and Archive**. Oak Ridge: Oak Ridge National Laboratory, September 2010. Disponível em: <http://daac.ornl.gov/PI/BestPractices-2010.pdf>. Acesso em: 10 out. 2015.

ICPSR. **Guide to Social Science Data Preparation and Archiving**. 2012. Disponível em: <http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf>. Acesso em: 10 out. 2015.

KNIGHT, Virginia; BOYD, David; GRAY, Stephen. **Research data management glossary**. Bristol: University of Bristol, 2013. Disponível em: <http://vocab.bris.ac.uk/data/glossary/>. Acesso em: 01 out. 2015.

MICHENER, William K. et al. Nongeospatial metadata for the ecological sciences. **Ecological Appliations**, v. 7, n. 1, p. 330-342, 1977. Disponível em: <http://lits.bio.ic.ac.uk:8080/litsproject/Micheneretal1997.pdf>. Acesso em: 10 out. 2015.

OECD. **OECD Principles and Guidelines for Access to Research Data from Public Funding.** 2007. Disponível em: **<**http://www.oecd.org/sti/sci-tech/38500813.pdf>. Acesso em: 01 out. 2015.

PEERJ. **Scientists who share data publicly receive more citations**. Disponível em: <http://www.eurekalert.org/pub_releases/2013-10/p-sws092413.php>. Acesso em: 10 out. 2015.

RESEARCH AND ENTERPRISE SERVICES. **Research data Management**. Disponível em: <https://research.ncl.ac.uk/rdm/glossary/>. Acesso em: 10 out. 2015.

RESEARCH DATA MANAGEMENT SERVICE GROUP. **Glossary of data management terms**. Disponível em: <http://data.research.cornell.edu/content/glossary>. Acesso em: 10 out. 2015

RESEARCH DATA OXFORD. **Research data Management Glossary**. Disponível em: <http://researchdata.ox.ac.uk/home/glossary/>. Acesso em: 10 out. 2015.

SAYÃO, Luís Fernando; SALES, Luana Farias. Dados abertos de pesquisa: ampliando os conceitos de acesso livre. **RECIIS – Rev. Eletron. de Comun. Inf. Inov. Saúde**. v. 8, n. 2, p. 76-92, 2014.

STRASSER, Carly et al. **Primer on Data Management**: What you always wanted to know. California: CDL, 2012. Disponível em: <http://escholarship.org/uc/item/7tf5q7n3#page-1>. Acesso em: 01 out. 2015.

SURA. **A Step-By-Step Guide to Data Management.** 2013. Disponível em: <http://www.lib.ua.edu/wiki/sura/index.php/A_Step-By-Step_Guide_to_Data_Management>. Acesso em: 01 out. 2015.

UK DATA ARCHIVE. **Craate & Manage Data**: formatting your data. Disponível em: <http://www.data-archive.ac.uk/create-manage/format>. Acesso em: 01 out. 2015.

WIGGINS, Andrea et al. **Data management guide for public participation in scientific research**. Albuquerque, NM: DataONE, 2013. Disponível em: <https://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf>. Acesso em: 10 out. 2015.

APÊNDICE I
GLOSSÁRIO DE TERMOS DE GESTÃO DE DADOS DE PESQUISA[60]

ANONIMIZAÇÃO DE DADO (*data anonimization*)	Processo pelo qual passam dados que contêm informações pessoais para que não revelem a identidade de indivíduos e evitem que esses dados não estejam vinculados a outros dados que possam revelar a identidade desses indivíduos.
ARQUIVAMENTO (*archiving*)	Serviço voltado para registrar, organizar e armazenar material digital com o objetivo de assegurar a longevidade e o acesso continuado a esses materiais. O serviço é baseado em aplicação de **METADADO**, políticas e metodologias de preservação digital e armazenamento seguro.
CADERNO DE CAMPO (*field notebook*)	Ferramenta usada por pesquisadores de várias áreas para fazer anotações quando executam trabalhos de campo. É um exemplo clássico de fonte primária. Esta ferramenta de pesquisa é geralmente usada por biólogos, geólogos, geógrafos, paleontólogos, arqueólogos, antropólogos (etnógrafos), etnomusicólogos e sociólogos[61].
CADERNO DE LABORATÓRIO (*laboratory notebook*)	Ferramenta usada por pesquisadores de várias áreas para fazer anotações sobre a pesquisa quando executada em laboratórios, criando um registro primário da pesquisa. Pesquisadores usam o caderno para documentar suas hipóteses, experimentos e análises iniciais ou interpretação desses experimentos; serve também como uma ferramenta de organização, memória e tem também um papel na proteção dos direitos de propriedade intelectual advindos da pesquisa[62].

[60] Com base nas seguintes fontes: a) STRASSER, Carly et al. **Primer on Data Management**: What you always wanted to know. California: CDL, 2012. Disponível em: <http://escholarship.org/uc/item/7tf5q7n3#page-1>. Acesso em: 01 out. 2015; b) KNIGHT, Virginia; BOYD, David; GRAY, Stephen. **Research data management glossary**. Bristol: University of Bristol, 2013. Disponível em: <http://vocab.bris.ac.uk/data/glossary/>. Acesso em: 01 out. 2015; c) RESEARCH DATA MANAGEMENT SERVICE GROUP. **Glossary of data management terms**. Disponível em: <http://data.research.cornell.edu/content/glossary>. Acesso em: 10 out. 2015; d) RESEARCH AND ENTERPRISE SERVICES. **Research data Management**. Disponível em: <https://research.ncl.ac.uk/rdm/glossary/>. Acesso em: 10 out. 2015; e) RESEARCH DATA OXFORD. **Research data Management Glossary**. Disponível em: <http://researchdata.ox.ac.uk/home/glossary/>. Acesso em: 10 out. 2015.

[61] Com base em <https://pt.wikipedia.org/wiki/Caderno_de_campo>

[62] Com base em <https://en.wikipedia.org/wiki/Lab_notebook>

CENTRO DE DADOS (*data center*)	Instalação equipada com computadores e dispositivos de armazenamento em massa usada para arquivamento e transmissão de dados. Os centros de dados frequentemente oferecem serviços de curadoria e gestão, acesso a produtos de dados, *help desk* e treinamento; em alguns casos oferecem também apoio a atividades de processamento de dados e outros serviços de valor agregado.
CIBERINFRAESTRUTURA (*cyberinfraestructure*)	Estrutura tecnológica que consiste de sistemas computacionais e armazenamento de dados, repositórios e ferramentas de computação que são ligados em rede, oferecendo recursos mais poderosos para a descoberta e a inovação.
CICLO DE VIDA DOS DADOS DE PESQUISA (*research data lifecycle*)	Operações que precisarão ser realizadas sobre os registros de dados de pesquisa durante toda a sua vida – desde o seu planejamento até o seu arquivamento ou descarte – para garantir que eles possam ter o seu uso, reuso e compartilhamento otimizado e estendido.
COLEÇÃO DE DADOS (*data set*)	Termo geral frequentemente usado para descrever um conjunto de dados de pesquisa que pode ser formado por um único elemento, como uma planilha de dados numéricos; pode igualmente ser formado por um conjunto de elementos relacionados, tais como planilhas, imagens, ou leituras diárias de um instrumento científico.
CONJUNTO DE DADOS	ver **COLEÇÃO DE DADOS**
CONTROLE DE QUALIDADE (*quality control*)	Atividades voltadas para identificar problemas nos dados. Veja também: **GARANTIA DE QUALIDADE**
COLEÇÕES DE DADOS DERIVADOS (*derived data set*)	Uma nova coleção de dados criada a partir de múltiplas coleções de dados já existentes e que usa os elementos de dados dessas coleções como fontes. Também se refere a uma nova coleção criada pela adição de uma única coleção de dados, usada como fonte, a novos dados coletados. Veja também: **DADOS DERIVADOS.**
CONSENTIMENTO INFORMADO (*informed consente*)	Processo de comunicação que permite que um indivíduo faça escolhas informadas sobre sua participação em uma pesquisa. Um acordo de consentimento informado fornece as informações necessárias sobre a pesquisa e serve como um compromisso formal para que uma pessoa participe voluntariamente de uma proposta de pesquisa.
CRIPTOGRAFIA (*cryttograpy*)	Codificação ou outra modificação sobre os dados com a finalidade de protegê-los de acesso e modificações não autorizadas, especialmente quando são transmitidos.

CURADORIA DE DADOS (*digital curation*)	Ações voltadas para o gerenciamento de dados de pesquisa durante o seu ciclo de vida; envolve manter, preservar e adicionar valor aos dados.
DADO ABERTO (*open data*)	Dados de pesquisa que são disponíveis livremente para reuso e republicação sem restrições de *copyright*, patentes ou outros mecanismos de controle de propriedade intelectual.
DADO BRUTO (*raw data*)	Dado que vem diretamente dos instrumentos científicos ou coletados diretamente da fonte sem sofrer nenhuma manipulação ou processamento. Também chamado de DADO CRU ou DADO PRIMÁRIO.
DADO CRU	Ver **DADO BRUTO**
DADO DERIVADO (*derivative data*)	Resultado do processamento ou combinação de **DADOS BRUTOS** ou de outros dados. Também chamado de DADO SECUNDÁRIO
DADO DE PESQUISA (*research data*)	Unidades de informação criadas ou coletadas no curso da pesquisa científica, e que são frequentemente formatadas de maneira a torná-las adequadas à comunicação, interpretação e processamento por computador. São exemplos de dados de pesquisa: planilhas de estatísticas, uma série de mensagens de e-mail, um registro sonoro de uma entrevista, um registro descritivo de um espécime de rocha, uma coleção de imagens digitais. Dependendo do contexto em que são consideradas, quase todas as coisas podem ser consideradas dados de pesquisa.
DADO PRIMÁRIO	Ver **DADO BRUTO**
DADO SECUNDÁRIO	Ver **DADO DERIVADO**
DOI - IDENTIFICADOR DE OBJETO DIGITAL (*digital object identifier*)	É um identificador persistente que é usualmente assinalado a itens digitais como um artigo de periódico ou uma coleção de dados, com o objetivo de identificá-los univocamente e dessa forma serem descobertos e citados.
GARANTIA DE QUALIDADE (*quality assurance*)	Conjunto de atividades direcionadas a assegurar que os dados são gerados e compilados de forma a atender os objetivos do projeto de pesquisa. Ver também: **CONTROLE DE QUALIDADE**.
GESTÃO DE DADOS DE PESQUISA (*research data management*)	Conjunto de práticas de gestão voltadas para o tratamento de dados de pesquisa durante o seu ciclo de

	vida; inclui todos os aspectos de manutenção, compartilhamento, segurança e preservação.
LIMPEZA DE DADOS (*data cleaning*)	Processo de eliminação ou edição de parte de dados que estão corrompidos ou sem a acurácia desejada, com o objetivo de alcançar o nível conveniente de integridades para a coleção de dados.
MELHORES PRÁTICAS (*best practices*)	Métodos ou enfoques que são reconhecidos por uma comunidade como sendo corretos ou mais apropriados para aquisição, gerenciamento, análise e compartilhamento de dados.
FLUXO DE TRABALHO CIENTÍFICO (*scientific workflow*)	Descrição precisa dos procedimentos científicos, frequentemente conceitualizados como uma série de dados.
FORMATO DE ARQUIVOS (*file format*)	Organização específica da informação em um arquivo digital.
INDICADOR DE NÍVEL DE QUALIDADE (*quality level flag*)	Indicador dentro do arquivo de dados que identifica o nível de qualidade de um dado particular.
META ANÁLISE (*meta-analysis*)	Análise que combina resultado de vários estudos.
METADADO (*metadata*)	Documentação ou informação sobre a coleção de dados; pode estar incorporado aos dados ou existir separadamente; metadados podem descrever, por exemplo, a autoria, direitos de propriedade, propósitos, métodos, organização e condições de uso dos dados, informações técnicas dos dados e outras informações necessárias à compreensão dos dados
PARÂMETRO (*parameter*)	Variável ou fator mensurável que determina ou caracteriza um sistema.
PERÍODO DE EMBARGO (*embargo period*)	Período de tempo na qual o acesso e uso dos dados para certos tipos usuários podem estar restrito, com o objetivo de proteger o interesse dos proprietários dos dados - pesquisadores e organização – e também de editores científicos.
PLANO DE GESTÃO DE DADOS DE PESQUISA (*research data management plan*)	Documento que formaliza o compromisso de como os dados que serão coletados ou gerados por um projeto de pesquisa, serão gerenciados e compartilhados durante o seu ciclo de vida.

PRESERVAÇÃO DE DADOS DE PESQUISA (*research data preservation*)	Conjunto de métodos tecnológicos e gerenciais voltados para garantir que os dados permaneçam intactos, acessíveis e compreensíveis ao longo do tempo.
PROVENIÊNCIA (*provenience*)	História do arquivo de dados ou da coleção de dados, incluindo a coleta, transformações, controle de qualidade, analises ou edição.
REPOSITÓRIO DE DADOS DE PESQUISA (research data repository)	Estrutura tecnológica e gerencial que permite que pesquisadores depositem seus dados de pesquisa para armazenamento e amplo acesso.
REUSO (*reuse*)	Uso dos dados para propósitos diferentes do qual eles foram coletados, geralmente por outros pesquisadores que não os autores dos dados.
VALOR AUSENTE (*missing value*)	Valor que não está no arquivo de dados porque a informação ou amostra não foi coletada, foi perdida, não foi analisada, é um valor impossível, etc. Um código específico indica que um valor está faltando e um indicador (*flag)* explicita a razão porque o valor está faltando.

APÊNDICE II
METADADOS PARA DADOS DE PESQUISA

PADRÕES GERAIS[63]

Padrões e ferramentas que não foram desenvolvidos especificamente para dados de pesquisa, mas que ao longo do tempo foram aplicados em várias disciplinas científicas.

CERIF - COMMON EUROPEAN RESEARCH INFORMATION FORMAT
<http://www.eurocris.org/cerif/main-features-cerif >

Padrão recomendado pela União Europeia para registrar informações sobre atividades de pesquisa. A partir da versão 1.6 vem incluindo recursos específicos para registro de metadados para coleção de dados.

DATACITE METADATA SCHEMA
<http://schema.datacite.org>

Lista de metadados mandatórios que devem ser registrados quando se assinala o DOI para uma coleção de dados. Os metadados são definidos para a identificação precisa e consistente com objetivo de apoiar a citação e a recuperação de coleção de dados.

DCAT - DATA CATALOG VOCABULARY
<http://www.w3.org/TR/vocab-dcat/>

É um vocabulário RDF projetado para facilitar a interoperabilidade entre catálogos de dados publicados na Web

DUBLIN CORE
<http://dublincore.org>

Um padrão neutro que pode ser aplicado a várias disciplinas e recursos, que pode ser facilmente compreendido e implementado. É um dos padrões de metadados mais conhecido e mais amplamente usada. Ele permite a composição de perfis de aplicação para áreas específicas, como exemplificado a seguir:

- **AGRIS Application Profile**
  <http://www.fao.org/docrep/008/ae909e/ae909e00.htm>
  Esquema de metadados criado para descrição, intercâmbio e recuperação de informações na área de agricultura.

- **Dryad Metadata Application Profile**
  <http://wiki.datadryad.org/Metadata_Profile>
  Um perfil de aplicação baseado no *Dublin Core Metadata Initiative Abstract Model,* usado para descrever dados multidisciplinares que estão subjacentes à literatura científica revisada por pares.

[63] DCC. **General Research Data**. Disponível em: <http://www.dcc.ac.uk/resources/subject-areas/general-research-data>. Acesso em: 10 out. 2015.

ORIENTADOS POR DISCIPLINA[64]

BIOCIÊNCIAS

ABCD - ACCESS TO BIOLOGICAL COLLECTION DATA
<http://wiki.tdwg.org/ABCD>

Padrão para acesso e intercâmbio de dados primários sobre biodiversidade, incluindo espécimes e observações.

DARWIN CORE
<http://rs.tdwg.org/dwc/index.htm>

Um corpo de padrões, incluindo um glossário de termos, que têm como objetivo de facilitar o compartilhamento de informações sobre a diversidade biológica por meio da disponibilização de definições de referência, exemplos e comentários.

EML - ECOLOGICAL METADATA LANGUAGE
<http://knb.ecoinformatics.org/software/eml/>

Especificação de metadados desenvolvida particularmente para disciplinas na área de Ecologia.

GENOME METADATA
<http://enews.patricbrc.org/faqs/genome-metadata-faqs/>

Dados descritivos sobre genoma no contexto do PATRIC (*Pathosystems Resource Integration Center*), consistindo de 61 campos de metadados que são organizados em sete grandes categorias: *Organism Info, Isolate Info, Host Info, Sequence Info, Phenotype Info, Project Info* e Outras.

CIÊNCIAS DA TERRA

AGMES - AGRICULTURAL METADATA ELEMENT SET
http://aims.fao.org/standards/agmes

Padrão semântico para descrição, descoberta de recursos, interoperabilidade e intercâmbio de dados para diferentes tipos de recursos informacionais na área de Agricultura.

AVM - ASTRONOMY VISUALIZATION METADATA
<http://www.virtualastronomy.org/avm_metadata.php>

Metadados para descoberta de recursos definidos de forma padronizada, voltados para a completa visualização de imagens astronômicas.

CIM - COMMON INFORMATION MODEL
<https://earthsystemcog.org/projects/es-doc-models/cim >

Modelo para descrever experimentos numéricos conduzidos pelo *Earth System Modelling Community*, incluindo o modelo que eles usam e os dados que eles produzem.

[64] DCC. **Disciplinary Metadata**. Disponível em: <http://www.dcc.ac.uk/resources/metadata-standards>. Acesso em: 10 out. 2015.

CIÊNCIAS EXATAS

CIF - CRYSTALLOGRAPHIC INFORMATION FRAMEWORK
<http://www.iucr.org/resources/cif>

Um padrão extensível de formato de arquivo e um conjunto de protocolos para o intercâmbio de dados cristalográficos e dados estruturais relacionados.

FITS - FLEXIBLE IMAGE TRANSPORT SYSTEM
<http://fits.gsfc.nasa.gov/fits_standard.html>

Usados pela comunidade de Astronomia para descrever originalmente imagens de telescópio, mas é agora uma família de padrões para descrever dados multidimensionais, incluindo dimensões espaciais, temporais e espectrais.

SDAC - STANDARD FOR DOCUMENTATION OF ASTRONOMICAL CATALOGUES
<http://cds.u-strasbg.fr/doc/catstd.htx>

Usado como uma alternativa para o FITS no arquivamento de dados astronômicos em uma forma mais acessível para seres humanos e ferramentas padronizadas de linhas de comando Unix.

CIÊNCIAS SOCIAIS & HUMANIDADES

DDI - DATA DOCUMENTATION INITIATIVE
<http://www.ddialliance.org/>

Padrão internacional amplamente usado para descrever dados das ciências sociais, comportamental e econômica. Expressadas em XML, as especificações dos metadados DDI dão suporte a todo o ciclo de vida dos dados de pesquisa.

QuDEx - QUALITATIVE DATA EXCHANGE FORMAT
<http://www.data-archive.ac.uk/create-manage/projects/qudex?index=1>

Modelo qualitativo de intercâmbio de dados para arquivamento e compartilhamento de dados.

SDMX - STATISTICAL DATA AND METADATA EXCHANGE
<http://sdmx.org>

Um conjunto de padrões técnicos e estatísticos e de diretrizes para serem usados no intercâmbio e no compartilhamento eficientes de dados e metadados estatísticos.

APÊNDICE III
ÍNDICE REMISSIVO DAS INTERROGAÇÕES SOBRE DADOS DE PESQUISA

COPYRIGHT