GESTÃO EDITORIAL DE PERIÓDICOS CIENTÍFICOS

tendências e boas práticas

Organizadores:
Lúcia da Silveira
Fabiano Couto Corrêa da Silva

PUBLICAÇÕES
UFSC - BIBLIOTECA UNIVERSITÁRIA

EDIÇÕES do BOSQUE
CFH - UFSC

Gestão Editorial de Periódicos Científicos:

tendências e boas práticas

Gestão Editorial de Periódicos Científicos:

tendências e boas práticas

Organizadores
Lúcia da Silveira
Fabiano Couto Côrrea da Silva

PUBLICAÇÕES
UFSC - BIBLIOTECA UNIVERSITÁRIA

EDIÇÕES do BOSQUE
CFH – UFSC

1ª edição | 2020

Organização
Lúcia da Silveira
Fabiano Couto Côrrea da Silva
Conselho Editorial - BU Publicações
Roberta Moraes de Bem
Andréa Figueiredo Leão Grants
José Paulo Speck Pereira
Luciana Bergamo Marques
Cristiano Motta Antunes
Comissão científica
Anna Khris Furtado Dutra
Anderson Mendes
Andréa Figueiredo Leão Grants
Clarissa Agostini Pereira
Gabriel Araldi Walter
Fabiano Couto Côrrea da Silva
Jorge Moisés Kroll do Prado
Juliana Aparecida Gulka
Lúcia da Silveira
Maria Bernardete Martins Alvez
Revisão ortográfica e gramatical
Zulma Neves de Amorim Borges

Normalização
Zulma Neves de Amorim Borges
Revisão Geral
Andréa Figueiredo Leão Grants
Juliana Aparecida Gulka
Lúcia da Silveira
Arte visual
Lara Benedet
Pablo Fiqueiredo
Cristiano Motta Antunes
Diagramação
Arnoldo Bublitz

BU Publicações UFSC
Campus Universtário Reitor João David F. Lima, Acesso Trindade.
Florianópolis, SC
conselhoeditorial.bu@contato.ufsc.br
+55 48 37219310
Edições do Bosque
Nuppe/CFH/UFSC
https://nuppe.ufsc.br

https://doi.org/10.5007/978-65-87206-08-0

Catalogação na fonte pela Biblioteca Universitária da Universidade Federal de Santa Catarina

G393

Gestão editorial de periódicos científicos [recurso eletrônico] : tendências e boas práticas / organizadores, Lúcia da Silveira, Fabiano Couto Côrrea da Silva. – 1. ed. – Florianópolis : BU Publicações/UFSC : Edições do Bosque/UFSC, 2020.
226 p. : il., gráf., tab.

ISBN 978-65-87206-08-0
E-book (PDF).

1. Periódicos eletrônicos. 2. Editores de periódicos. 3. Tecnologia – Serviços de informação. 4. Ciência da Informação. I. Silveira, Lúcia. II. Silva, Fabiano Couto Côrrea da.

CDU 001:655.52

Elaborada pelos bibliotecários Jonathas Troglio e Suélen Andrade – CRB-14/1093 e CRB-14/1666

https://doi.org/10.5007/978-65-87206-08-0/2

Capítulo 2

Gestão de dados científicos para periódicos

Fabiano Couto Corrêa da Silva

Planta dente-de-leão em preto e raízes em branco. Dente-de-leão apresenta mais ramificações, a flor inicia a abrir as primeiras pétalas.

O que você vai encontrar neste capítulo:

- ✓ *Apresentação dos benefícios da gestão de dados científicos;*
- ✓ *Orientação para o editor das necessidades de recomendar aos autores que façam a gestão de dados científicos;*
- ✓ *Exemplificação de como qualificar os dados científicos.*

1 INTRODUÇÃO

O esforço realizado em pesquisas geralmente é reconhecido somente pelos resultados comprovados de teorias formuladas com embasamento em dados científicos, ainda que sejam poucos os pesquisadores realmente preocupados com o registro dos seus dados, descartando muitos deles após a publicação dos resultados. Frente à necessidade de estruturar o crescente volume de dados científicos, no processo de gestão a confiabilidade e a facilidade de uso dos dados são fundamentais, mas, em geral, ainda é um desafio implementar os requisitos de gestão necessários.

A exposição de tudo aquilo que foi utilizado no trabalho de pesquisa pode mostrar os erros e incertezas que o pesquisador não descreveu nas suas publicações. A transparência no processo de criação do conhecimento científico, o seu acesso e preservação são aspectos que devem ser analisados pelos envolvidos na produção e uso da informação científica de todas as áreas. Os planos de gestão de dados permitem não apenas a comprovação necessária para a avaliação dos resultados de uma pesquisa, mas também garantem que os objetivos dos trabalhos sejam compatíveis com seu acesso e preservação.

Para que sejam úteis à comunidade científica, porém, os dados devem seguir uma estrutura e organização clara, e constituir coleções informativas relacionadas e registradas em um formato adequado ao tema tratado, isto é, no contexto de uma determinada comunicação científica. Dessa forma, dos resultados gerados em uma pesquisa obter-se-á um conjunto de dados que poderá armazenar e ser reutilizado ao distribuir-se a outros pesquisadores, e inclusive poderá ampliar-se às áreas distantes às dos objetivos iniciais da pesquisa.

Mesmo que existam muitas atividades relacionadas com o uso de dados científicos, há um grande desconhecimento sobre como realizar esse objetivo,

e se já existem iniciativas nesse sentido. É comum que os editores de periódicos deparem-se com necessidades frequentes no cotidiano das suas atividades, incluindo demandas do tipo "Necessito que os revisores de um artigo da minha revista acessem um *conjunto de dados*"; "Necessito urgentemente um identificador persistente para um *conjunto de dados*"; "Necessito indexar *conjunto de dados* com metadados completos e em acesso aberto"; "Necessito atribuir licenças para o *conjunto de dados*", dentre muitas outras carências. Essas demandas para a gestão de dados científicos inclui processos de preservação, uso e reutilização que devem ser assimilados pelos pesquisadores e, principalmente, editores de revistas científicas, tendo em vista o papel que desempenham junto à cadeia produtiva da comunicação científica. Dominar esses aspectos é fundamental para que os editores de periódicos ofereçam condições para os pesquisadores arquivarem os dados que sustentam os resultados e argumentos das suas pesquisas.

Diante dessa perspectiva, analisaremos o processo de registro dos dados científicos e o papel dos editores de periódicos. Para isso, devem oferecer uma ampla gama de políticas de gestão a suas respectivas comunidades de pesquisa, incluindo o acesso a catálogos e bases de dados através de internet, protocolos de retenção, criação de metadados, migração de dados através de *software* e sistemas de *hardware*, e a formação e o desenvolvimento das normas internacionais. Ao oferecer esses serviços, os editores desempenham um papel ativo e estratégico na formulação de novos métodos e técnicas para intercâmbios de dados e adoção de novos padrões em todos os aspectos relacionados com sua conservação.

2 DE QUAIS DADOS ESTAMOS FALANDO?

Referimo-nos aqui a fatos, medidas, gravações, registros ou observações sobre o mundo, coletados por cientistas e outros, com um mínimo de interpretação contextual. Os dados podem estar em qualquer formato ou meio tomando a forma de notas, números, símbolos, imagens, filmes, vídeos, gravações sonoras, reproduções pictóricas, desenhos ou outras representações gráficas, manuais de procedimentos, formulários, diagramas, trabalhos fluxogramas, descrições de equipamentos, arquivos de dados, algoritmos de processamento de dados, registros estatísticos, etc.

O conceito de dados científicos também faz referência às distintas ferramentas, como protocolos, códigos numéricos, gráficos e tabelas que são necessárias para recolher e organizar os dados, tanto em trabalhos de campo quanto em laboratório. Incluem não somente os materiais e amostras biológicas ou ambientais extraídas, mas também os resumos gerados durante o transcurso da realização de uma pesquisa (SILVA, 2017). Todo conteúdo digital e não digital tem o potencial de tornar-se dado científico. Os dados científicos podem ser dados experimentais, dados observacionais, dados operacionais, dados de terceiros, dados do setor público, dados de monitoramento, dados processados ou dados adaptados.

Os dados científicos são a evidência que sustenta a resposta à pergunta de pesquisa e podem ser usados para validar os resultados, independentemente de sua forma (por exemplo, impressa, digital ou física). Estas podem ser informações quantitativas ou declarações qualitativas coletadas por pesquisadores no decorrer de seu trabalho por experimentação, observação, modelagem, entrevista ou outros métodos, ou informações derivadas de evidências existentes. Os dados podem ser brutos ou primários (por exemplo, diretos de medição ou coleta) ou derivados de dados primários para análise ou interpretação subsequente (por exemplo, limpos ou extraídos de um conjunto maior de dados) ou derivados de fontes existentes em que os direitos podem ser mantidos por outros.

Para serem localizáveis ou detectáveis, os dados e metadados devem ser descritos detalhadamente para permitir a pesquisa baseada em atributos. Para ser amplamente acessível, dados e metadados devem ser recuperáveis em uma variedade de formatos que são sensíveis a seres humanos e máquinas usando identificadores persistentes. Para ser interoperável, a descrição dos elementos de metadados deve seguir as diretrizes da comunidade, que usam um vocabulário aberto e bem definido. Para ser reutilizável, a descrição dos elementos de metadados essenciais, recomendados e opcionais deve ser processável por máquina e verificável. O uso deve ser fácil, e os dados devem ser citáveis para sustentar o seu compartilhamento e poder reconhecer-se o valor que possuem.

Os requisitos para compartilhar dados científicos são bastante recentes, e os sistemas para sua coleta e gestão ainda se encontram em processo de desenvolvimento. Por isso, na atualidade, a procura de um conjunto de dados concreto não é tão fácil como a de um artigo publicado, mesmo que a previsão seja a melhoria em um futuro próximo. Portanto, há algumas estratégias que podem auxiliar na procura de dados científicos.

Para encontrar um conjunto de dados, é recomendável começar a busca em artigos sobre o tema de interesse. Normalmente, os dados são depositados como material complementar de um artigo ou vinculados a eles.

Se a localização dos dados não depende de artigo, há algumas alternativas. A primeira consiste em buscar no currículo do autor para verificar se há alguma referência sobre a disponibilidade dos dados em alguma parte. Se isso não funciona, também é possível contatar diretamente o autor para solicitar-lhe o acesso a seus dados. As políticas de algumas revistas e organismos financiadores exigem uma cópia dos dados, sempre e quando os dados não sejam sensíveis. Nenhuma dessas estratégias é infalível, pois os dados mais antigos perdem-se e as direções de correio eletrônico mudam, porém, pode ser uma boa estratégia para obter acesso aos dados que correspondam a um artigo.

Se a procura está direcionada para os dados gerais de um tema e não para os dados de um artigo específico, a estratégia de busca será diferente. Um bom lugar para começar a procura de um tema é um índice das matérias específicas de uma especialidade, sempre que exista. Por exemplo, o Integrated Ocean Observing System (IOOS) enumera uma ampla gama de recursos marinhos e conta com um portal de busca para ajudar a encontrar os dados específicos das pesquisas sobre oceanos. Esses índices não necessariamente recolhem dados, mas apontam uma série de recursos sobre um tema em particular, conjuntamente com bases de dados que também possam estar disponíveis nas bibliotecas.

Na ausência de uma base de dados ou de uma biblioteca, também pode-se considerar a procura nos repositórios de dados que são populares em um determinado campo e que podem ser localizados na lista re3data.2. Deve-se levar em consideração também as fontes externas de dados, como agências governamentais, fundações de pesquisa, grupos de interesses especiais e outras organizações, pois frequentemente fazem com que os dados relacionados com suas atividades tornem-se disponíveis. Por exemplo, a Administração Oceânica e Atmosférica Nacional (NOAA) dos Estados Unidos é um excelente recurso para tudo o que se refere a dados relacionados com o clima. Com qualquer outro tipo de informação, sempre é recomendável avaliar a fonte de dados para assegurar-se sobre a credibilidade de sua obtenção.

Por último, sabemos que a medida que o intercâmbio de dados torne-se mais habitual, também será mais fácil encontrar dados com fins de reutilização. O

processo de pesquisa atualmente está em transição para um regime de intercâmbio de dados, o que significa que muitos de seus sistemas de intercâmbio e reutilização estão em vias de desenvolvimento para que, no futuro, seja tão simples encontrar os dados de um artigo como é agora encontrar o próprio artigo.

3 UM BREVE PANORAMA

A atividade científica gera continuamente dados científicos e, embora seja incentivada a publicação dos resultados de pesquisas em periódicos de acesso aberto, no Brasil a maioria dos dados ainda é publicada de forma incipiente (SILVA, 2017).

A tendência desejada pelas agências de fomento e pela sociedade é que a comunidade científica compartilhe os dados resultantes de suas pesquisas para serem reutilizados por outros pesquisadores. Porém, a realidade mostra que muitas revistas não possibilitam que os autores coloquem o resultado de sua atividade científica ao alcance de todos.

Os pesquisadores têm sido historicamente relutantes em compartilhar seus avanços científicos e os resultados da suas pesquisas, entre outras razões, por causa do medo de que seus pares reutilizem dados científicos de forma fraudulenta, sem serem reconhecidos por seu trabalho. Para combater esse problema, foram estabelecidas regras e diretrizes de conduta (ALLEA, 2011; DATA SHARING FOR THE PREVENTION OF FRAUD, 2011), e licenças que cobrem essas necessidades (Creative Commons, Open Data Commons, etc.), ajudando a mostrar uma atitude mais aberta com relação à disseminação de suas descobertas. No entanto, a relutância dos pesquisadores continua, seja por falta de informações sobre os procedimentos que devem adotar, ou mesmo a falta de infraestrutura adequada.

Por outro lado, embora os avanços ocorridos nos últimos anos tenham se dissolvido, em certa medida, também encontramos barreiras tecnológicas relacionadas com a falta de infraestrutura para armazenar dados corretamente e questões de padronização no formato dos dados. O W3Consortium recomenda a utilização de formatos específicos para o compartilhamento de dados, embora existam grandes quantidades de dados em formato eletrônico que não foram tratados (dados brutos), dificultando ou impedindo a sua utilização (BERNERS-LEE, 2019).

Nosso objetivo é analisar os recursos e políticas editoriais das revistas e repositórios e como elas afetam o depósito, autoarquivamento e reutilização, com relação

ao material suplementar (dados científicos). Concentraremos o estudo em diferenciar essas publicações que permitem o armazenamento e reutilização de dados abertos, não aqueles que aceitam o livre acesso aos postos de trabalho, sem especificar qual tratamento é destinado para os dados científicos. Daí a importância em diferenciar previamente os termos Acesso Aberto (Open Access) e Dados Abertos (Open Data):

O termo Acesso Aberto é definido como o acesso à literatura científica, disponível gratuitamente através da Internet, permitindo a leitura, *download*, cópia, distribuição, impressão, busca ou vínculo, por meio de *links,* ao texto completo dos artigos coletados para indexação. Tudo isso para fins legítimos, sem barreiras legais ou econômicas, permitindo assim o acesso através da internet para todos. A única restrição de reprodução e distribuição deve ser dada pelos autores com o controle sobre a integridade de seu trabalho e o direito de ser apropriadamente reconhecido e citado (BERLIN DECLARATION ON OPEN ACCESS TO KNOWLEDGE IN THE SCIENCES AND HUMANITIES, 2003). Existem duas maneiras para seguir os princípios de acesso aberto:

- Publicação em periódicos de acesso aberto, que é chamada de rota dourada (*gold road*);
- Arquivamento de trabalhos científicos em repositórios (institucionais ou temáticos), que é chamado de rota verde (*green road*).

O termo Open Data é o movimento que promove a liberação de dados, geralmente não textuais, em formatos reutilizáveis como o CSV[1, AI]ém disso, o Open Data Handbook (2019) define dados abertos como aqueles que podem ser reutilizados e distribuídos livremente por qualquer pessoa, sujeitos ao requisito de atribuição de autoria e reutilização da mesma forma em que aparecem.

1 CSV (os valores separados por vírgulas) são um tipo de formato aberto de documento simples para a representação de dados em forma de tabela, na qual as colunas estão separadas por vírgulas.
O formato CSV é muito simples e não indica um conjunto de caracteres específico, nem como os ***bytes*** estão localizados nem o formato da quebra de linha. Esses pontos devem ser indicados muitas vezes ao abrir o arquivo, por exemplo, com uma planilha. O formato CSV não é padronizado. A idéia básica de separar campos com uma vírgula é muito clara, mas torna-se complicada quando o valor do campo também contém aspas duplas ou quebras de linha. As implementações de CSV podem não manipular esses dados ou usar citações de outro tipo para envolver o campo. Mas isso não resolve o problema: alguns campos também precisam incorporar essas citações, portanto, as implementações de CSV podem incluir caracteres ou seqüências de escape.

As vias existentes para publicar os dados científicos são as seguintes: repositórios (institucionais, temáticos e dados); sites institucionais ou pessoais; revistas de dados; material suplementar em artigos de periódicos.

Atualmente, existem bancos de dados e projetos que identificam quais são as políticas relacionadas a direitos autorais, condições de reutilização e autoarquivamento dos principais editores de revistas especializadas, embora se refiram ao acesso aberto à publicação, não ao material. Por exemplo, o diretório internacional SHERPA/RoMEO, o Registry of Open Access Repositories (ROAR) e o Directory of Open Access Repositories(OpenDOAR) e o DOAJ. No diretório ROARMAP, há uma lista completa de repositórios, mandatos e políticas relacionadas a acesso e dados abertos, que inclui mais de 800 políticas internacionais, organizadas de acordo com sua origem (agência de financiamento, provedor de financiamento ou organizações de pesquisa). Todos são iniciativas que incluem a localização, o tipo de acesso ao seu conteúdo, as políticas de direitos autorais e as condições de arquivamento das publicações, bem como o nível de adesão dos editores com relação ao autoarquivamento. Diferenciam entre a versão preliminar de um artigo que ainda não foi publicado (pré-impressão) e o artigo final publicado em uma revista (pós-impressão). Como indicamos, eles são genéricos e não especificam a política que rege os dados científicos; referem-se apenas ao acesso aberto das publicações e não analisam particularmente o material suplementar.

Outra iniciativa muito relevante no cenário internacional é o ODiSEA (International registry on Research Data), um diretório internacional dos periódicos que admitem dados científicos. Atua com a coleta de periódicos que aceitam material suplementar, examinando as políticas de direitos autorais de editores científicos para identificar arquivos digitais que contêm dados científicos, repositórios de dados, etc.

ODiSEA é dividido de acordo com as áreas de conhecimento do Essential Science Indicators Web of Knowledge: Agronomia, Biologia e Química, Química, Medicina Clínica, Ciência da Computação, Economia e Negócios, Engenharia, Ecologia Ambiente, Geociências, Imunologia, Ciência dos Materiais, Matemática , Microbiologia, Biologia Molecular e Genética, Multidisciplinar, Neurociências e Comportamento, Farmacologia e Toxicologia, Física, Planta e Zootecnia, Psiquiatria / Psicologia, Ciência Social Geral, Ciência Espacial.

Existem iniciativas que incentivam o acesso de dados científicos abertos, tanto nacional como internacionalmente. Um deles é o Horizon 2020, um programa de

pesquisa e inovação da União Europeia para o período 2014-2020. O seu orçamento é de aproximadamente 80 milhões de euros e está empenhado em promover a excelência científica e a liderança industrial na comunidade europeia, desenvolvendo a ciência a partir de conhecimentos anteriores. Nas suas orientações para o acesso aberto (EUROPEAN COMMISSION, 2017), são explicadas as formas recomendadas para alcançar a máxima divulgação científica por meio de publicações em acesso aberto (OA Ouro e OA Verde) e de dados científicos.

A União Europeia está empenhada em melhorar o acesso à informação científica e aumentar os benefícios do investimento público, com a premissa de que não deve ser necessário pagar por informações que tenham sido financiadas com fundos públicos, garantindo que elas facilitem o acesso à informação. Nessa perspectiva, entendem que os editores de periódicos científicos devem assegurar que todas as publicações revisadas por pares possam ser lidas, baixadas e impressas. E incentivem, na medida do possível, direitos adicionais, como o direito de copiar e distribuir o material. Nesse sentido, foi criado o Open Research Data Pilot (ORD Pilot), concebido para maximizar o acesso e a reutilização de dados científicos, sendo possível:

a) Depositar os dados científicos necessários para validar publicações em um repositório de dados aberto, juntamente com seus metadados. Embora não seja obrigatório depositar todos os dados gerados durante a investigação, apenas aqueles que são essenciais, "tão abertos quanto possível, tão fechados quanto necessário" (EUROPEAN COMMISSION, 2017).
b) Eles devem adotar medidas para permitir o acesso a terceiros, com o uso de licenças como a Creative Commons (CC BY ou CCo)[2].

Após definir o repositório dos dados, é recomendável atribuir uma licença para que eles sejam reutilizados e distribuídos livremente por qualquer pessoa, permitindo a cópia, distribuição, transmissão, *download* e criação de trabalhos de-

2 As licenças Creative Commons não substituem os direitos autorais, mas dependem delas para permitir que os termos e condições de uma licença de trabalho sejam escolhidos da maneira mais adequada para o detentor dos direitos. Por esse motivo, essas licenças têm sido entendidas por muitos como uma maneira pela qual os autores podem assumir o controle de como desejam compartilhar sua propriedade intelectual.

rivados. No capítulo 3, foram discutidos os tipos de licenças para uso não comercial do Creative Commons que são geralmente usadas nesse contexto.

4 INFRAESTRUTURAS OPEN SCIENCE

São repositórios que qualquer pesquisador pode usar, independentemente de sua filiação institucional, para preservar qualquer tipo de produção acadêmica. Os dois exemplos mais conhecidos são *Figshare* e Zenodo.

Para conseguir seu objetivo de abertura dos dados, a Comissão Europeia colocou à disposição o repositório de dados abertos Zenodo. Criado por OpenAIRE e CERTh com o apoio da Comissão Europeia, esse repositório oferece seus serviços a partir da iniciativa pan-europeia OpenAIRE, que amplia a vinculação dos resultados da pesquisa com a informação sobre *Conjunto de dados* e financiamento em contextos europeus e nacionais.

Zenodo é uma iniciativa do portal OpenAire, que dispõe de uma infraestrutura adequada para a hospedagem de *conjunto de dados* e outros resultados de pesquisa de projetos europeus. Está construído sobre a plataforma Invenio e desenvolvimento no CERN, centro que se ocupa também da gestão da enorme quantidade de dados do Large Hadron Collider (LHC). Como no caso de *Figshare,* o acesso ao depósito é livre, atribuído DOI e permite conjuntos de dados disponíveis em BibTeX, EndNote e outros formatos bibliográficos. Os usuários podem agregar metadados a seus arquivos, muito mais detalhados que em *Figshare*, um espaço próprio utilizando metadados sob licença CC0, ou seja, dedicadas ao domínio público sem restrições nem solicitações de permissões, exceto para endereços de *e-mail*. Além disso, sempre que permitido, outros usuários Zenodo podem comentar seus arquivos, e uma interessante característica é que precisa que seja fácil inscrever-se com seu identificador ORCID ou conta de GitHub.

Figshare é uma plataforma criada por *Digital Science* que permite compartilhar e mostrar os resultados de pesquisas multidisciplinares e que está dirigida a pesquisadores, cientistas, projetos e instituições. Atualmente está associada com F1000 *Research* (um prestigioso repositório de artigos científicos), colabora com *PLOS* (a maior revista científica de acesso aberto do mundo) e também com *Plum Analytics* (um serviço que quantifica o impacto dos trabalhos de pesquisa

publicados). Todo o material publicado em *Fisgshare* é identificado com um DOI para facilitar sua localização e sua citação. Na plataforma, podemos localizar: apresentações, vídeos, pôsteres, imagens, dados, artigos, etc., e a preservação dos dados funciona com tecnologia CLOCKSS, uma organização sem fins lucrativos que promove a aliança entre os editores do mundo acadêmico e as bibliotecas acadêmicas para arquivar, de um modo sustentável, todo o conteúdo *web* produzido no âmbito científico.

Os usuários podem integrar os dados do repositório com outros *websites* e *blogs,* copiando e colando um simples código. Os leitores podem realizar comentários sobre os conjuntos de dados e fazer *download* em arquivos de citação a seus gestores de referência para seu uso posterior. O repositório também oferece a possibilidade de publicar resultados negativos ou sobre experimentos fracassados para que outros pesquisadores poupem o esforço de terem de passar por testes já realizados e, dessa forma, não percam muitas horas de trabalho em determinados casos.

Caso não haja repositórios públicos estruturados na sua área de conhecimento, incluindo a possibilidade de uma alternativa aos repositórios Figshare e Zenodo, há outras alternativas possíveis, como o Dataverse, DataHub, DANS, Mendeley Data, Google Dataset Search e o Repositório Multidisciplinar Dryad, que seguem os Princípios FAIR (Findable, Accessible, Interoperable, Reusable).

5 BOAS PRÁTICAS PARA SALVAR CONJUNTO DE DADOS

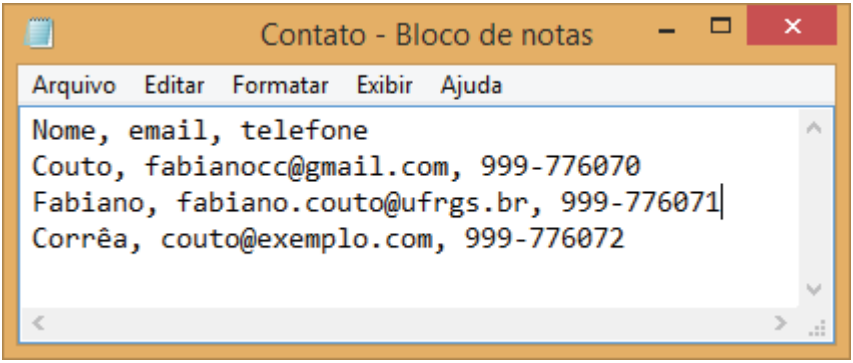
Conjunto de dados é a principal representação dos processos de análise de dados, são demonstrados em "formato de planilha em que as linhas são os registros dos acontecimentos, e as colunas são as características desses acontecimentos." (AQUARELA, 2018, p. 1). Para que um conjunto de dados seja organizado adequadamente, é necessário ser apresentado por valores atribuídos que caracterizem seu real conteúdo.

No formato de arquivo de planilha, os dados são armazenados nas células. Cada célula é organizada em linhas e colunas. Uma coluna no arquivo de planilha pode ter tipos diferentes. Por exemplo, uma coluna pode ser do tipo sequência, um tipo de data ou um número inteiro. Alguns dos formatos de arquivo de planilha mais populares são valores separados por vírgula, no formato .CSV (comma separated values), Planilha do Microsoft Excel (xls) e Planilha XML aberta do Microsoft Excel (xlsx).

Cada linha no arquivo CSV representa uma observação, normalmente chamada de registro. Cada registro pode conter um ou mais campos separados por vírgula.

Às vezes, é possível encontrar arquivos em que os campos não são separados por vírgula, mas separados por tabulação (Figura 1). Esse formato de arquivo é conhecido como formato de arquivo TSV (valores separados por tabulações).

Figura 1 - Arquivo CSV aberto no bloco de notas



Fonte: Elaboração do autor – captura de tela (2019).
Descrição da imagem: Captura de tela do software bloco de notas com conteúdo de nome, e-mail e telefone separados entre vírgulas, demonstrando que essa separação, ao dar-se em outro sistema, transforma-se em linhas e colunas contínuas com a Tabela 1. Fim da descrição.

Arquivos no formato CSV oferecem a possibilidade de organizar ilimitada quantidade de informações para conjuntos de dados com uma entrada por linha e campos separados por vírgulas (ou outros separadores). É por isso que é útil como um formato simples de troca de dados que pode ser importado, manipulado e exportado por várias aplicações, principalmente processadores de planilhas, formulários xml, etc. (JORGE; DOUGLAS, 2016). Um exemplo simples da aparência de um arquivo CSV, pode ser realizado com o seguinte exercicio, ao abrir um *software* editor de texto e insita as seguintes informações:

Nome	**Idade**
Fabiano,	41
Barreto ,	60

Salve como CSV e abra o arquivo pelo Excel, os dados são transferidos para as células e são divididos pela vírgula, conforme tabela 1:

Tabela 1 – transferência dos dados para planilha

Nome	Idade
Fabiano	41
Barreto	60

Fonte: Elaboração do autor.

Mesmo criando um arquivo CSV em outro programa ou software de banco de dados, também é possível abri-lo no Excel como uma pasta de trabalho usando o comando "Abrir" no menu "Arquivo". No entanto, existem algumas considerações especiais que é necessário levar em consideração ao abri-lo.

Abrir um arquivo CSV no Excel não altera o formato do arquivo, apenas permite que você abra seu arquivo CSV no programa e visualize o conteúdo. Há outras etapas necessárias para criar a conversão completa:

Etapa 1 - Abra o Microsoft Excel
Etapa 2 - Clique em "Arquivo" e toque em "Abrir".
Etapa 3 - Será aberta a caixa de diálogo Abrir. Quando abrir essa janela, selecione "Arquivos de texto (* .prn, *. Txt, *. Csv) na lista suspensa. Essa lista estará visível no canto inferior direito da caixa.
Etapa 4 - Navegue por seus arquivos para localizar o arquivo CSV. Abra-o clicando duas vezes no arquivo.

Na sequência, o Excel tentará abrir o arquivo imediatamente. No caso de um arquivo .csv, o Excel deve abri-lo e importar os dados em uma nova pasta de trabalho sem problemas.

Se estiver abrindo um arquivo .txt, precisará passar pelo Assistente de importação de texto para que os dados sejam localizados corretamente.

Quando o Microsoft Excel abre o arquivo .csv, ele usa todos os dados padrão das configurações para entender como importar o arquivo e atribuir dados a cada uma das colunas da pasta de trabalho. Se o conjunto de dados tiver for-

matação diferente dos padrões do programa, será necessário usar o Assistente para Importação de Texto.

6 A QUALIDADE DOS CONJUNTO DE DADOS

O desenvolvimento da pesquisa científica atualmente pautada em coleções de dados torna necessário que sejam acessíveis e rastreáveis da mesma maneira que as publicações tradicionais. Portanto, a qualidade também desempenha um papel importante em ambos os tipos de produtos de pesquisa. Porém, enquanto nas publicações esse aspecto foi operacionalizado - nem sempre sem controvérsia - por meio de revisão por pares e índices de citação, ele ainda está engatinhando no que diz respeito aos dados científicos.

Existem vários propósitos para a retenção e o compartilhamento de conjuntos de dados além do período ativo inicial de captura e análise, dentre eles:

1. apoiar a revisão por pares de publicações com base nos dados;
2. permitir a validação dos resultados;
3. compartilhar com a próxima geração de pesquisadores ou usuários que executam trabalhos semelhantes;
4. reunir-se com uma comunidade mais ampla de usuários;
5. apoiar a preservação e o acesso a longo prazo para conjuntos de dados selecionados como valor de longo prazo.

É importante que todos os conjuntos de dados sejam representados em uma estrutura que revele as características de itens de dados individuais e os relacionamentos entre eles. Um formato de conjunto de dados adequado para preservação deve manter a integridade sintática da estrutura e dos valores individuais, para que a análise automatizada seja possível. Também é essencial para a usabilidade dos dados o entendimento da semântica dos elementos de dados e seus relacionamentos no conjunto de dados. A semântica pode ser descrita explicitamente dentro do conjunto de dados, descrita explicitamente em um documento auxiliar (de preferência processável por máquina) ou implícita através da conformidade com as melhores práticas da comunidade ou especificação externa.

Em campos com uma base de conhecimento da comunidade acumulada, o desafio pode ser menos a preservação de conjuntos de dados individuais do que a migração de um sistema inteiro para novas tecnologias. Para algumas classes de conjuntos de dados, a característica mais significativa para futuros usuários é a capacidade de integrar conjuntos de dados individuais nos sistemas de informações atuais e futuros. Isso é fundamental em campos em que os dados são de uso intermitente contínuo, como dados relacionados à política, medicina, etc.

7 INCONSISTENCIAS NOS METADADOS

Ao cadastrar um conjunto de dados, é recomendável diferenciar o título do artigo correspondente, uma vez que a recuperação pode associar ambos os arquivos como se tivessem o mesmo conteúdo. Devido à sua natureza de embasamento de uma pesquisa, os dados científicos podem ser reaproveitados em novas investigações e, naturalmente, seu conteúdo ser relacionado para novas pesquisas.

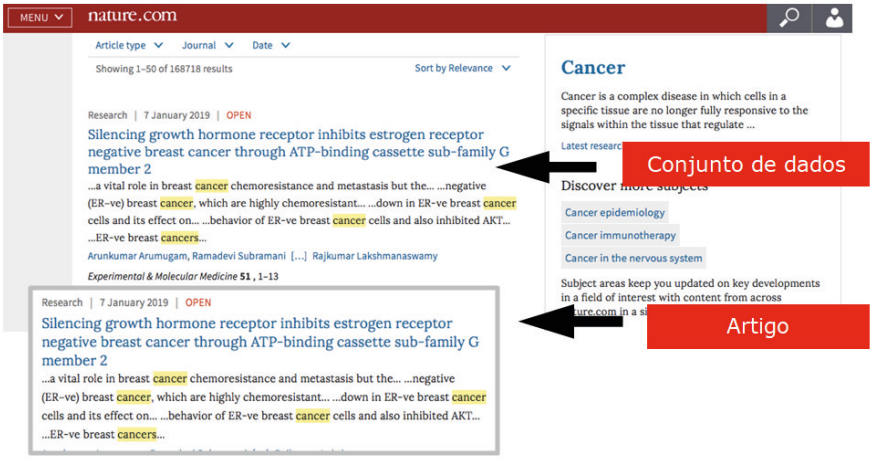
Por exemplo, observamos na revista PhytoKeys[3] o artigo indicando o Zenodo[4] para o conjunto de dados, no entanto, foi encontrado algumas fragilidades: 1) apresenta títulos iguais nos metadados. 2) conteúdo incompleto: Enquanto o artigo demonstra um mapa, acompanhado de uma tabela o outro apresenta apenas um mapa em relação a unidade ou conjunto de dados representado no repositório Zenodo.

Quando é atribuído o mesmo título do artigo e ao conjunto de dados (Figura 3), pode acarretar confusão na sua busca, ou seja, o usuário que estiver realizando o levantamento de determinados dados encontrará duplicidade de documentos como se ambos carregassem o mesmo conteúdo: conjunto de dados e artigo.

Para a completude e maior abrangência da descrição dos metadados é recomendável que seja incorporado diferentes idiomas para representar os conjuntos de dados, assim, terá maior possibilidade de serem recuperados.

3 Disponível em: https://zenodo.org/record/1138141#.XY5PX-hKjIU
4 Disponível em: https://zenodo.org/record/1138143#.XY5WCuhKjIU

Figura 3 - Conjunto de dados com o mesmo título



Fonte: Nature (2019).

Descrição da imagem: *Site* do Mega-Journal *Nature* com cabeçalho vermelho e branco, apresenta os metadados do artigo *Silencing growth hormone receptor inhibits estrogen receptor negative breast cancer throught ATP-bingding cassette sub-family G member* 2. Compara duas telas: uma com os dados, e a outra do artigo; ambas possuem os mesmos metadados. Duas setas apontam para esses metadados, indicando que um é o artigo, e o outro é o conjunto de dados. Fim da descrição.

7 CITAÇÃO DE ACORDO COM OS PRINCÍPIOS FORCE11

A *Joint Declaration of Data Citation Principles* foi criada pelo grupo de trabalho internacional FORCE11, constituído por uma comunidade de pesquisadores, bibliotecários, arquivistas, editores e agências de fomento de pesquisa científica, todos interessados no avanço da comunicação científica. Essa declaração foi assinada por mais de 80 das principais editoras científicas, universidades e instituições do mundo, entre as quais destacamos Elsevier, PLoS, ORCID, Nature Publishing Group, Association of Research Libraries, BioMed Central, CrossRef, etc. Os objetivos da iniciativa são conseguir que, uma vez que se estabeleça a cultura de citação dos dados, comecem a ser evidentes os benefícios; entre eles:

1. A infraestrutura editorial deve assegurar que as referências eletrônicas aos dados mantenham-se no futuro e possam ser reutilizadas;
2. Os serviços de publicação eletrônica deverão construir controles para que diminua o perigo de que pesquisadores "roubem" dados alheios (plágio de dados);
3. O impacto, tanto dos conjuntos de dados como o dos criadores desses dados, poderá ser medido; Os autores beneficiar-se-ão com a atribuição e o crédito no trabalho facilitados pela citação. Isso ajuda a citar a propriedade intelectual apropriadamente, facilitando sua busca e, consequentemente, seu impacto;
4. Os pesquisadores poderão obter reconhecimento profissional da mesma maneira que obtêm pelas publicações tradicionais.
5. Para os leitores, tornar-se-á mais fácil pesquisar e encontrar conjuntos de dados quando eles têm uma citação formal;
6. Os componentes da referência bibliográfia dos conjuntos de dados são:
   autores, ano; título do conjunto de dados; Data do arquivamento no repositório, versão (se houver); identificador persistente (exemplo DOI).

Veja o exemplo aplicado:

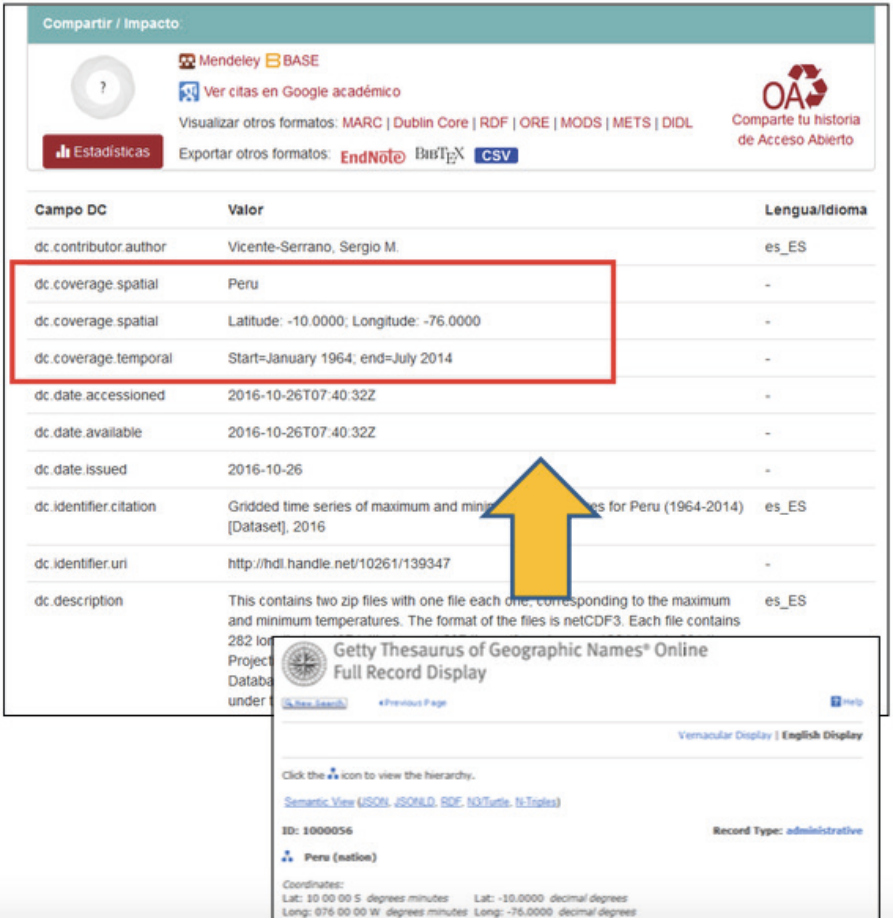
VICENTE-SERRANO, Sergio M. 2016. Gridded time series of maximum and minimum temperatures for Peru (1964-2014), [Dataset], DIGITAL.CSIC, http://dx.doi.org/10.20350/digitalCSIC/7389

7.1 Informação cronológica e geográfica nas referências:

a) Dc.coverage.temporal: refere-se às datas em que a referência de dados / temporal foi coletada, na forma start = XXXX; end = XXXX.

b) Dc.coverage.spatial: refere-se ao local onde os dados foram coletados / referenciados, a melhor prática é usar formulários padronizados (Getty Thesaurus de Nomes Geográficos, GEONAMES) e incluir coordenadas de latitude e longitude. A Figura 4, apresenta os campos descritores dos metadados do conjunto de dados no Mendeley data, incluindo os dados de localização.

Figura 4 - Informação geográfica e cronológica



Fonte:Captura de tela do Mendeley Data (2019).

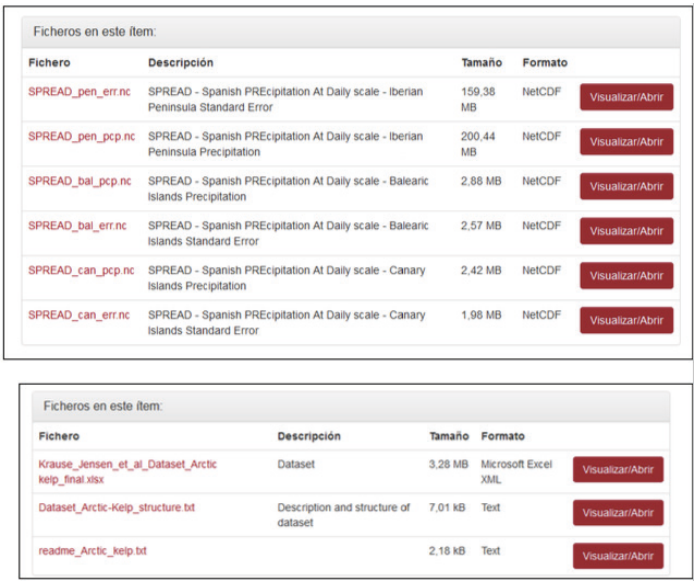
Nesse caso apontado na figura 4 a referência ficou sem os dados de localização de acordo com o Mendeley data, mas o mais adequado seriam estar presente.

11 METADADOS SOBRE *SOFTWARES* E FORMATOS

Uma boa prática é oferecer os dados em vários formatos, aqueles que são mais usados em uma disciplina específica e, em seguida, em um formato aberto. É recomendável indicar a versão do conjunto de dados, no título e na referência bibliográfica, além de indicar as alterações na descrição.

Também é recomendável indicar nos metadados se algum *software* é necessário para abrir e usar os dados, preferencialmente indicando onde acessar o *software* (Figura 5).

Figura 5 - Metadados sobre *softwares*, formatos e readme



Fonte: Captura de tela do Mendeley Data (2019).
Descrição de imagem: a tela lista o conjunto de dados relacionados a uma mesma pesquisa. São quatro colunas com as seguintes categorias: título do artigo, descrição, tamanho, formato e botão de visualizar o item. O destaque foi para o formato do arquivo, que nesse caso são: NetCDF, exel, text e também para o arquivo readme comentado na próxima seção. Fim da descrição.

12 ARQUIVO README (LEIA-ME) SIGNIFICATIVO

Fornece informações sobre o conjunto de dados para que seja interpretado corretamente por pessoas e máquinas, deve-se elaborar um *readme file* (leia-me) por conjunto de dados, ou quando o item é composto de vários conjuntos de dados.

O título do arquivo leia-me deve ser nomeado de maneira que possa ser facilmente associado ao conjunto de dados; em formato plano p.e txt.

O conteúdo desse arquivo precisa descrever brevemente o conjunto de dados; o contato do investigador principal; a data da coleta de dados, a data de criação do conjunto de dados; a informação geográfica dos dados; a metodologia, o *link* para publicações e outras documentações relacionadas a esse conjunto de dados; a unidades de medida, os protocolos, as abreviaturas, os códigos, os símblos associados aos dados; a licença de uso; e a citação recomendada. A seguri apresenta-se o conteúdo do readme do conjunto de dados tratados anteriormente na Figura 4 e 5.

```
This dataset includes 5 km. spatial resolution time
series of maximum and minimum temperatures for the
entire Peru. The gridded data has been created using
the entire temperature series available for Peru,
which were subjected to a quality control and ho-
mogenization procedure. Gridded data was created by
means of a regression-based approach using terrain
and topographic variables as inputs. One independent
model was created for each month of the series. Resi-
duals were interpolated by means of a IDW procedure.
The data was validated using a jackknife approach.
Details of the methodology and validation results can
be found at: Vicente-Serrano, S.M., Juan I. Lopez-Mo-
reno, Kris Correa, Grinia Avalos, Juan Bazo, Cesar
Azorin-Molina, Fernando Domínguez-Castro, Ahmed El
Kenawy, Luis Gimeno, Raquel Nieto, Recent changes in
monthly surface air temperature over Peru, 1964-2014.
Submitted to International Journal of Climatology.

The dataset contains two zip files with one file each
one, corresponding to the maximum and minimum tempe-
ratures.

The format of the files is netCDF3.

Each file contains 282 longitudes, 407 latitudes and
607 times (from January 1964 to July 2014).
```

```
Projection is Geographic (WGS84). The mean monthly
temperature is in °C.

Contact person: Sergio M. Vicente Serrano svicen@ipe.
csic.es

Issue date: 26 October, 2016

Identifiers: https://digital.csic.es/han-
dle/10261/139347, http://dx.doi.org/10.20350/digi-
talCSIC/7389.

Access and reuse: http://opendatacommons.org/li-
censes/odbl/1-0/.
```

13 CONSIDERAÇÕES FINAIS

É consenso entre a comunidade científica que os dados devam ser preservados para garantir o seu acesso no futuro. Mas há um grande desconhecimento sobre como realizar esse objetivo e se já existem iniciativas nesse sentido.

O gerenciamento de dados científicos inclui processos de preservação, uso e reutilização. Dominar esses aspectos é fundamental para que os pesquisadores planejem seu trabalho desde a concepção do projeto até a execução, uso e arquivamento. O apoio dos editores de periódicos aos pesquisadores tem como objetivo auxiliar na obtenção de recursos, em nível conceitual e prático, entre o processo de pesquisa e o depósito de dados, tendo em conta as infraestruturas disponíveis e as possibilidades de utilização dos recursos nas diferentes áreas do conhecimento. No presente capítulo, apresentamos recomendações sobre o uso apropriado de recursos tanto para armazenamento quanto à disseminação de dados. Para incentivar o aprendizado sobre algumas das tecnologias existentes para a preservação de dados, faremos exercícios práticos de indexação dos dados com o uso de ferramentas abordadas. É necessário identificar a demanda de cada situação em particular, mas é recomendável que as escolhas sempre sejam direcionadas para recursos que fomentem a ciência aberta, possibilitando um ciclo de retroalimentação contínuo que beneficie toda a cadeia da produção científica,

pesquisadores e sociedade.

REFERÊNCIAS

ALLEA. **The european code of conduct for research integrity.** 2011. Disponível em: https://ec.europa.eu/research/participants/data/ ref/h2020/other/hi/h2020-ethics_code-of-conduct_en.pdf. Acesso em: 9 jan. 2019.

BERNERS-LEE, T. **W3C Standards**. 2019. Disponível em: https://www.w3.org/standards. Acesso em: 1 jan. 2019.

EUROPEAN COMMISSION. **Guidelines to the rules on open access to scientific publications and open access to research data in horizon 2020**. 2017. Disponível em: http://ec.europa.eu/research/ participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. Acesso em: 8 jan. 2019.

JORGE, L. F. F.; DOUGLAS, F.. Avaliação de formatos de publicação de dados abertos governamentais através de indicadores de usabilidade. Tendências da Pesquisa Braileira em Ciência da Informação; João Pessoa Tomo 9, n. 1, 2016.

MAX PLANCK SOCIETY; MAX PLANCK INSTITUTE FOR THE HISTORY OF SCIENCE. **Berlin declaration on open access to knowledge in the sciences and humanities**. Berlín: Max Planck, 2003.

MENDELEY DATA. 2019. Disponível em: https://data.mendeley.com. Acesso em: 23 jul. 2019.

NATURE. 2019. Disponível em: https://www.nature.com. Acesso em: 23 jul. 2019.

OPEN DATA HANDBOOK. 2019. Disponível em: http://opendatahandbook.org. Acesso em: 8 jan. 2019.

PHYTO KEYS. 2019. Disponível em: https://phytokeys.pensoft.net. Acesso em: 23 jul. 2019.

PUNDIR, S. **FAIR data principles**. 2019. Disponível em: https://commons.wikimedia.org/wiki/ File:FAIR_data_principles.jpg. Acesso em: 1 jan. 2019.

SILVA, F. C. C. da. **Gestão de dados científicos**. Rio de Janeiro: Interciência, 2019.

SILVA, F. C. C. da. Gestión de datos oceanográficos: propuesta de un modelo para Brasil. 2017. 260 f. **Tese.** Doctorado en Información y Documentación en la Sociedad del Conocimiento. Facultad de Biblioteconomia, Universitat de Barcelona, Barcelona, 2017.

WILKINSON, M. *et al.* The FAIR guiding principles for scientific data management and stewardship. **Sci Data**, [s*. l.*], n. 3, 2016.

ZENODO. 2019. Disponível em: https://zenodo.org. Acesso em: 23 jul. 2019.