DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333

ARTIGO

O PAPEL DOS BIBLIOTECÁRIOS NA GESTÃO DE DADOS CIENTÍFICOS

THE ROLE OF LIBRARIANS IN SCIENTIFIC DATA MANAGEMENT

EL PAPEL DE LOS BIBLIOTECARIOS EN LA GESTIÓN DE DATOS DE INVESTIGACIÓN

Fabiano Couto Corrêa[1]

RESUMO: Apresentamos uma análise das possibilidades de atuação do bibliotecário em conjunto com pesquisadores para um gerenciamento eficiente de dados científicos. O resultado obtido, por meio de um levantamento de ferramentas e técnicas de gestão atualmente disponíveis, demonstra um quadro analítico de ações de apoio que os bibliotecários podem fornecer para a elaboração de um projeto para o ciclo de vida dos dados científicos. Conclui-se que a gestão de dados científicos exigem soluções de planejamento que incluem conhecimentos específicos sobre a escolha do repositório e técnicas de armazenamento para a conservação e o uso permanente dos dados como chave para o êxito de um projeto de pesquisa.

PALAVRAS-CHAVE: Dados científicos. Ciclo de vida dos dados. Preservação digital. Bibliotecas universitárias. Repositórios científicos. E-ciência. Curadoria de dados. Acesso livre. Bibliotecários acadêmicos.

ABSTRACT: We present an analysis of the possibilities of the librarian profession with researchers for efficient management of scientific data. The result obtained through a survey of tools and currently available management techniques, demonstrates an analytical framework of support actions that librarians can provide for the development of a project on the life cycle of scientific data. It is concluded that the scientific data management requires planning solutions that include specific knowledge about the repository choice and storage techniques for the conservation and the permanent use of the data as a key to the success of a research project.

KEYWORDS: Scientific data. Data life cycle. Digital preservation. University Libraries. Scientific Repositories. E-science. Data curation. Free access. Academic librarians.

RESUMEN

Se presenta un análisis de las posibilidades de acción del bibliotecario en conjunto con los investigadores para una gestión eficiente de los datos científicos. El resultado obtenido a través de un análisis de herramientas y técnicas de gestión disponibles en la actualidad, demuestra un marco analítico de acciones de apoyo que los bibliotecarios pueden ofrecer para el desarrollo de un proyecto para el ciclo de vida de los datos científicos. Se concluye que la gestión de datos científicos requieren soluciones de planificación que incluyen conocimientos específicos sobre la elección de las técnicas de depósito y almacenamiento para la conservación y el uso permanente de los datos como una clave para el éxito de un proyecto de investigación.

[1] Doutorando em Información y Documentación en la Sociedad del Conocimiento (Universidad de Barcelona), Professor do Instituto de Ciências Humanas e da Informação (ICHI) da Universidade Federal do Rio Grande (FURG). Porto Alegre, RS. Email: fabianocc@gmail.com - ORCID: http://orcid.org/0000-0001-5014-8853.
Submetido em: 03/08/2016 – **Aceito em**: 30/08/2016.

PALABRAS CLAVE: Datos científicos.Ciclo de vida de los datos. Preservación digital. Bibliotecas universitarias. Repositorios científicos. E-ciencia. Curaduría de datos. Acceso libre. bibliotecarios académicos.

1 INTRODUÇÃO

A preservação e intercâmbio de dados científicos converteram-se em temas de interesse em escala internacional para gestores, agências financiadoras e investigadores em geral. Com isso, atualmente, a demanda de apoio para a gestão dos dados científicos requer que os bibliotecários compreendam e respondam novas demandas dos pesquisadores, não somente como consumidores de informação, mas também como produtores.

Diante deste novo cenário no fluxo da comunicação científica os bibliotecários estão auxiliando os pesquisadores em um nível mais amplo do processo de pesquisa, em vez de se concentrar unicamente em meios formais de comunicação científica.

Do ponto de vista do acesso aos dados da investigação, as bibliotecas estão desenvolvendo serviços de apoio durante as fases do ciclo de vida dos dados científicos, ou seja, quando os pesquisadores estão gerando e utilizando o dados em seu plano de trabalho. Muitas vezes, estes serviços devem ser prestados em estreita colaboração com pesquisadores e podem incluir o desenvolvimento de planos de gestão para documentar e organizar os dados através do desenvolvimento de ferramentas ou recursos para armazenar dados de forma segura.

Um dos principais problemas de gestão de dados científicos está relacionado com a falta de continuidade do registo de dados decorrente da saída de algum membro do projeto ainda em andamento; por isso é normal que surjam dificuldades de reutilizar um conjunto de dados, uma vez que, sem a devida documentação, torna-se difícil compreender como, quando e por que os dados foram capturados.

Uma eficiente gestão de dados reduz a quantidade de trabalho necessária para a interpretação e compilação de informações obtidas no final de um projeto de pesquisa quando se encontra bem documentado, assim a investigação em curso não precisará ser reconstruída em uma data posterior. Por isso, a assistência do bibliotecário aos pesquisadores ajuda a avaliar o que realmente necessitam entender, a compreenderem como organizar uma variedade de tipos de dados e tomar as decisões corretas sobre o acesso e preservação de dados para os seus projetos.

A seguir relacionaremos o conjunto de ações que o bibliotecário pode oferecer para a comunidade científica em seus projetos.

1.2 Etapa 1

A primeira fase é a obtenção de dados. Neste momento, os pesquisadores devem encontrar uma maneira eficiente de armazenar os dados coletados, decidir sobre uma

DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333

estratégia de metadados e assegurar que todos os participantes da pesquisa compreendem o esquema de metadados.

Há uma variedade de formas de apoio em que o bibliotecário pode se envolver, desde o desenvolvimento de um plano de gestão ou estratégia de metadados, e também escolhendo um repositório adequado para as necessidades dos investigadores.

Após os dados terem sido criados ou recolhidos, os investigadores começam a processá-los. Isso envolve a transcrição, digitalização, validação, limpeza e armazenamento de dados durante o processo de registro. Os bibliotecários podem oferecer soluções de armazenamento ou ferramentas que ajudam os pesquisadores a gerar os metadados durante a investigação. Durante a fase de análise, os pesquisadores interpretam os dados e desenvolvem a investigação.

1.3 Etapa 2

A próxima etapa é a preservação de dados, que envolve a migração para formatos adequados para a preservação, incluindo a criação de backups e gerando metadados adicionais. Nesta fase, os bibliotecários podem ajudar os pesquisadores que não conhecem os requisitos para a preservação de dados, fornecendo informações sobre formatos de preservação, ou podem migrar dados diretamente. Também é possível oferecer oficinas de formação para os pesquisadores sobre o plano de gerenciamento de dados que pode ser projetado para integrar um projecto de investigação desde o início.

É importante considerar que muitos repositórios apresentam embargos para acessar os dados. Por exemplo, quando os dados forem armazenados num repositório também é possível determinar quando os *datasets* estarão disponíveis publicamente. Esta opção é útil quando um pesquisador quer que os dados sejam preservados, mas não estão preparados para torná-los disponíveis para consulta pública. Além disso, a maioria dos repositórios de dados estão em desenvolvimento e pode ser difícil distinguir o mais adequado para as necessidades específicas de cada investigador, de modo que o uso e a verificação das possibilidades de cada um deve ser um processo contínuo. A primeira questão para perguntar sobre um repositório é quem é o responsável; por exemplo, o financiador da pesquisa ou universidade. Para um pesquisador ou grupo de pesquisa que se desenvolve a longo prazo, será importante que o repositório seja bem gerenciado ao longo do tempo.

A segunda pergunta é quanto tempo os dados devem ser armazenados. Esta questão nem sempre é explicitamente respondida, mas se o depósito é baseado em LOCKSS (Lots of Copies Keep Stuff Safe), CLOCKSS (Controlled LOCKSS) o Portico é um bom sinal. Estes serviços asseguram que os dados permaneçam disponíveis mesmo que o repositório não esteja. Outra questão é de que maneira o repositório indexa os conjuntos de dados, ou seja, os formatos de arquivos e a padronizarão de metadados.

Depois de escolher o repositório, o pesquisador precisa fornecer informações sobre a forma como os dados serão utilizados e os requisitos de acesso, as restrições e informações que permitam avaliar a qualidade dos dados. É muito importante que a implementação de um repositório de dados seja um projeto em que o ciclo de vida envolva decisões conjuntas entre o pesquisador e o bibliotecário, de modo que as necessidades e as possibilidades técnicas sejam abordadas no plano de gestão. Todos esses fatores ajudam a garantir que os conjuntos de dados sigam com um formato utilizável, assim como controles que facilitem sua busca em outros repositórios. Finalmente, nunca é demais perguntar para outros pesquisadores e bibliotecários o que recomendam para a preservação de dados em seu campo de atuação.

Se um pesquisador deseja compartilhar dados de um artigo publicado, o primeiro lugar que deve consultar é a revista que publicou o artigo. Um número crescente de revistas exigem o intercâmbio de dados e recomenda especificamente onde os dados devem ser depositados para facilitar o processo de revisão por pares. Por exemplo, a revista Scientific Data tem uma lista de recomendações repositórios para os autores (Scientific Data, 2015). Algumas revistas facilitam a inclusão dos seus dados em um repositório próprio, como é o caso das revistas integradas com o repositório Dryad. Obviamente, nem todas as revistas têm estas recomendações repositório, mas é aconselhável seguir estas orientações, sempre que existam.

Um segundo aspecto a considerar é onde os investigadores de uma determinada área de pesquisa compartem seus dados. A seleção de um repositório reconhecido pela comunidade científica de uma determinada área torna os dados mais propensos a serem descoberto por pesquisadores de áreas afins e muito mais provável de serem citados.

Também é recomendável buscar opções locais para o intercâmbio de dados através de um repositório institucional. Se deve ter em conta que algumas universidades participam em consórcios de repositórios de dados, tal como a 3TU.Datacentrum (2015), com o apoio da Universidade de Tecnologia de Eindhoven, Universidade de Tecnologia de Delft e da Universidade de Twente. Se um repositório institucional está disponível, vale a pena considerar como um lugar para depositar os dados, mesmo que seja apenas para melhorar a assistência local. Como a gestão de dados e intercâmbio estão se tornando cada vez mais proeminente, muitas instituições já oferecem suporte para o processo de preservação dos dados científicos.

DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333

2 PLANEJAMENTO DE GESTÃO DE DADOS

Gradualmente o planejamento de gestão de dados científicos esta se tornando um requisito dos financiadores e instituições que pagam para coleta de dados, uma vez que demonstram preocupação com a disponibilidade dos dados no futuro, como se pode verificar as chamadas, por exemplo, o programa Horizonte 2020, em que a Comissão Europeia lançou um projeto piloto chamado "Open Research Data Pilot" para promover e otimizar a gestão e reutilização de dados de pesquisa gerados pelos projetos que financia.

A coleta, organização e preservação da informação tem sido confiada a bibliotecários de diferentes setores que trabalham com os fluxos de informação; a articulação dos planos de gestão de dados é simplesmente uma manifestação moderna destas funções. Uma maneira de oferecer suporte simultaneamente a vários pesquisadores é oferecendo *workshops* sobre o desenvolvimento de planos de gestão de dados. Além disso, é possível oferecer consultas individuais para a elaboração dos planos de gestão de dados, geralmente realizando entrevistas sobre o conteúdo que necessitam desenvolver. Muitas vezes, existem problemas de preservação e documentação que os investigadores nunca consideraram e o bibliotecário, por meio dessas entrevistas, podem alertá-los.

Em geral, os pesquisadores não tem certeza se seguem corretamente regras para o depósito e organização dos dados científicos, também não estão bem informados sobre a propriedade intelectual e apresentam dúvidas sobre as exigências para a elaboração de planos de gestão de dados. Os investigadores devem garantir que as normas institucionais estão sendo seguidas e que todas as decisões se ajustam em conformidade com estas regras ou com as necessidades específicas do projeto.

Em conclusão, os planos de gestão de dados podem melhorar através da colaboração entre bibliotecários, profissionais, pesquisadores e repositório, pois todos apresentam diferentes experiências e conhecimento para ajudar na criação de um plano de gerenciamento de dados. Os bibliotecários, por exemplo, pode estabelecer uma autenticação institucional e personalizar a ferramenta com recursos adicionais (texto de ajuda, sugestões de respostas, etc.).

3 ENTREVISTAS COM OS PESQUISADORES

Embora a preservação tenha significado muito específico para bibliotecários, pesquisadores apresentam conceitos muito diferentes sobre o processo de arquivamento. Os bibliotecários devem começar a discussão sobre a preservação assegurando que as partes interessadas tenham um entendimento comum dos termos.

A entrevista é o momento certo para identificar os requisitos do plano de gestão de dados. Pesquisadores da Universidade de Purdue (Carlson, 2012) realizaram um estudo que indica que os bibliotecários possuem habilidades específicas para realizar entrevistas que podem ser úteis para o desenvolvimento de planos de gestão de dados, principalmente a

capacidade de negociar e gerenciar as expectativas usuários. Salientam que os bibliotecários necessitam conhecer, antes de realizar a primeira entrevista com os investigadores, os recursos disponíveis e os tipos de pesquisa que está sendo conduzida em suas instituições, os tipos de recursos disponíveis nos departamentos do campus e práticas de pesquisa mais frequentes. Isso não significa que eles devem ter o conhecimento de um perito, porque, em muitos casos, o bibliotecário aprenderá mais durante o trabalho conjunto com os pesquisadores. As habilidades mais importantes são o conhecimento sobre quais perguntas devem ser realizadas e onde encontrar as respostas.

De acordo com o estudo da Universidade de Purdue, uma boa ferramenta para começar a pensar sobre como trabalhar com pesquisadores é definir um perfil de curadores de dados. Os pesquisadores da Purdue University Libraries e da Escola de Graduação em Biblioteconomia e Ciência da Informação da Universidade de Illinois desenvolveram um conjunto de perfis de curadores de dados que descreve como os pesquisadores criam e gerenciam dados primários. Em seguida, organizaram uma oficina e criaram um conjunto de ferramentas para ajudar os bibliotecários a entender o que eles precisavam saber antes de realizar uma entrevista. O produto deste trabalho, o Data Curation Profiles Toolkit, inclui uma planilha para professores e um manual de entrevistas para bibliotecários. Os autores do manual estabeleceram uma série de questões simples e fáceis de entender, buscando uma linguagem comum entre ambos. As perguntas permitem que bibliotecário e pesquisador analisem o ciclo de vida dos dados, como as informações são compartilhadas durante o projeto e como o acesso deve ser fornecido tanto durante a investigação quanto após o seu término. A planilha também pode ser útil para que o pesquisador possa descrever todos os problemas relacionados aos dados e visualizá-los em um único lugar.

Um projeto de pesquisa semelhante foi realizado na Universidade de Colorado. Lage e Maness (2011) realizaram extensas entrevistas com pesquisadores e desenvolveram oito perfis que representam os docentes e estudantes de pós-graduação que foram entrevistados. Estes perfis revelam que as necessidades, práticas e compreensão da investigação e os dados variam muito entre as diferentes disciplinas. E, embora cada instituição, departamento e pesquisador exibam qualidades únicas, esses perfis podem ser úteis para entender os problemas que surgem quando se trabalha com os investigadores. Os perfis desenhados pela equipe da Universidade do Colorado descrevem o nível de interesse, o apoio que os pesquisadores sentem que recebem, problemas de armazenamento que enfrentam e a privacidade que requerem os dados. Um fator que a equipe observou é que nem sempre existe um retorno positivo entre o nível de apoio que os pesquisadores necessitam e a participação do bibliotecário. Algumas entrevistas correlacionaram positivamente a receptividade de um pesquisador com a participação de bibliotecários no gerenciamento de dados. Também evidenciou a falta de suporte para armazenamento e preservação de dados, um viés positivo em direção ao movimento de acesso aberto e falta de suporte para o gerenciamento de dados durante o processo de investigação. A equipe de Colorado também observou que os pesquisadores da área de ciências da terra pareciam mais abertos à participação da biblioteca

e para disponibilizar seus dados. Aqueles que trabalham em domínios altamente competitivos, como as ciências exatas, demonstraram ser menos receptivos à participação da biblioteca no processo de gestão de dados.

Embora Lage (2011) e seus colegas tenham descoberto uma vasta gama de atitudes em relação à partilha de dados e curadoria, surgiram alguns pontos em comum entre muitos dos pesquisadores entrevistados. A maioria deles não identificou os dados das suas pesquisas como dados públicos, mas isso não significa necessariamente que eles não estavam abertos para compartilhar, já que os pesquisadores, muitas vezes, buscam manter um certo nível de controle sobre os dados que compartilham.

O estudo realizado pela Universidade do Colorado também mostra que os pesquisadores concordam com os procedimentos ou serviços departamentais de armazenamento de dados, que a maioria dos pesquisadores têm alguns subconjuntos de dados de pesquisa que não estão sendo mantidos ou conservados com um plano em desenvolvimento e que percebem as tarefas de gestão de dados como distrações de seus projetos de pesquisa. Essas atitudes revelam que é importante que o gerenciamento de dados seja menos complicada quanto possível para que os pesquisadores e que o planejamento deve ser considerado antes de iniciar a coleta de dados.

Depois de identificar os pesquisadores receptivos para trabalhar com a biblioteca e que solicitem ajuda com o plano de gerenciamento de dados, chega o momento de realizar entrevistas especificamente sobre os seus projetos de pesquisa. A entrevista pode ser guiada por um esboço de um plano de gestão ou um modelo do ciclo de vida dos dados.

O bibliotecário pode começar perguntando se o pesquisador conhece as ferramentas que podem ser usadas para preservar dados e ajudar a identificar formatos de arquivos apropriados para depósito e planos de armazenamento de dados.

Após estas questões é recomendável continuar com perguntas sobre segurança. Estar ciente das opções para realizar backup dos dados disponíveis para os pesquisadores é útil neste caso, como o conhecimento das vantagens e desvantagens das diferentes opções de armazenamento. Os dados podem ser armazenados em discos rígidos locais internos ou externos, em sistemas baseados em nuvem ou em servidores. Para obter mais informações sobre opções de armazenamento, é aconselhável consultar a unidade de tecnologia da informação sempre que possível. Garritano e Carlson (2009) sugerem definir o fluxo de trabalho incluindo o processamento e os passos analíticos realizados após a coleta de dados. Para levar adiante uma boa gestão de dados, não se trata apenas de selecionar uma boa opção de armazenamento, mas também políticas, melhores práticas e suporte para backup e armazenamento.

Muitas vezes, os pesquisadores mantem os dados armazenados depois que o projeto está concluído no mesmo lugar onde eles foram armazenados durante o processo de investigação, independentemente de como o procedimento adotado poderá afetar a

capacidade de uso acessibilidade a longo prazo. Arquivar dados implica a preservação ativa de dados, bem como a adoção de medidas para aumentar a capacidade de serem encontrados e acessibilidade. Se trata, entre outras coisas, de registrar códigos identificadores únicos para os dados e para a realização de controles comuns para sua replicação.

Um repositório de dados funciona mediante processos de indexação que agregam valor ao conteúdo disponível ao invés de simplesmente guardá-lo. É importante que os pesquisadores entendam a diferença entre o armazenamento simples e arquivar corretamente. Se os pesquisadores armazenam os dados digitais em servidores ou discos rígidos sem executar regularmente as ações de preservação necessárias, com o tempo seus dados se tornarão inutilizáveis. Também é necessário saber quem possui e controla os dados e se existe problemas de privacidade. Em muitos casos, os pesquisadores não têm respostas fáceis para essas perguntas, de modo que os bibliotecários podem ajudá-los a ir além das respostas imediatas. Os bibliotecários possuem habilidades únicas para organizar informações dispares com a finalidade de desenvolver um projeto de pesquisa e para encontrar fontes de referência como manuales, que podem ajudar os pesquisadores a compreender melhor as ferramentas que podem ser úteis, o contexto da investigação e o formato necessário para indexação dos dados. Também é aconselhável para falar com o diretor do laboratório do projeto ou responsável pelos instrumentos de coleta de dados: ele pode ajudar a entender como os dados foram gerados ou recolhidos. Se não houver um investigador principal do projeto, pode ser útil discutir este fluxo de trabalho com as pessoas que estão reunindo dados diretamente. O bibliotecário também pode ajudar os pesquisadores a fazer conexões com outras pessoas da mesma comunidade disciplina, as quais podem fornecer suporte com ferramentas e todas as informações necessárias sobre o tema em questão. Este é o ponto onde as habilidades do bibliotecário e sua capacidade de encontrar informações são úteis, porque; apesar de não saber todos os repositórios ou padrões de metadados, o bibliotecário pode encontrar as respostas, graças ao seu conhecimento universal de fontes de informação. Uma vez que os pesquisadores estejam familiarizados com os pontos que entram na gestão de dados, desde a planificação e gerenciamento de metadados para preservação a longo prazo, serão mais conscientes dos componentes que poderiam deixar de fora ou incluir no processo de planejamento para garantir que a preservação seja eficaz.

Além disso, uma entrevista poderia ser uma conversa ou uma série de consultas para ensinar os recursos de que necessitam através do ciclo de vida dos dados de pesquisa. É provável que o armazenamento de dados já seja parte do trabalho de um pesquisador, mas os métodos devem ser explicitamente discutidos e acordados no início de um projeto. O fluxo de trabalho flui corretamente se cada colaborador de um projeto de pesquisa utiliza os mesmos métodos de armazenamento de dados e está familiarizado com eles.

DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333

4 SELEÇÃO DE REPOSITÓRIOS

No processo de seleção de repositórios bibliotecários podem ajudar a compreender as características fundamentais de cada um; por exemplo, repositórios temáticos proporcionam visibilidade dentro da comunidade correspondente, na medida em que a opção para o repositório institucional pode ser mais eficiente para dar visibilidade ao desenvolvimento de pesquisas com os pares mais próximos (por exemplo, do mesmo instituto, em uma universidade).

Os editores podem aceitar conjuntos de dados associados com artigos; o bibliotecário também pode recomendar o depósito em um repositório de terceiros ou de apoio à publicação repositórios bibliotecas podem optar por aceitar conjuntos de dados em um repositório institucional ou um tanque especialmente projetado especificamente para um conjunto de dados para uma investigação específica.

Existem várias opções disponíveis para depositar dados científicos, como por exemplo o re3data[2], que apresenta um extenso diretório de repositórios em todas as disciplinas. A seleção mais adequada deve ser feita no início de um projeto de pesquisa; também pode ser apropriado armazenar dados em mais de um repositório. Por exemplo, todos os dados de projecto podem ser armazenadas num repositório institucional, enquanto subconjuntos de dados também podem ser armazenados em repositórios de domínio específico para aumentar as possibilidades de ser localizado. O bibliotecário pode ajudar pesquisadores neste processo de decisão, pesquisando as opções disponíveis e orientá-los para fazer uma avaliação significativa dos seus dados. A seleção do repositório deve levar em conta questões como a existência de um repositório específico para uma disciplina apropriada ou de um repositório temático para um conjunto de dados, as políticas de acesso para o repositório e as políticas de de preservação.

O processo de envio de dados no repositório selecionado pode ser muito variável. Pesquisadores devem consultar os requisitos do repositório visando a preparação de dados, para o qual a maioria tem um guia passo a passo com instruções detalhadas sobre os formatos e outras questões técnicas. Às vezes, assume a forma de um sistema de envio por e-mail; por isso os pesquisadores podem submeter os seus dados via e-mail e os responsáveis pelo repositório realizam o trabalho restante. Assim, os bibliotecários podem prestar assistência aos investigadores, com o depósito de dados ou a oferta de instruções com um passo a passo para repositórios comuns com o auxílio de guias da biblioteca.

Eventualmente os repositórios são muito específicos e não apresentam links para publicações e outros conjuntos de dados que lhes dão contexto. Alguns podem ter pacotes de dados inativos devido à falta de continuidade de uma investigação ou por outras razões. Por

[2] Registry of Research Data Repositories. Disponível em: www.re3data.org

DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333

isso, é aconselhável verificar se o tipo de dados do repositório em que estamos interessados é atualizado regularmente.

Para entender melhor a variedade de repositórios existentes, estabelecemos cinco categorias que descrevem a seguir, incluindo alguns exemplos.

4.1 Repositórios Institucionais

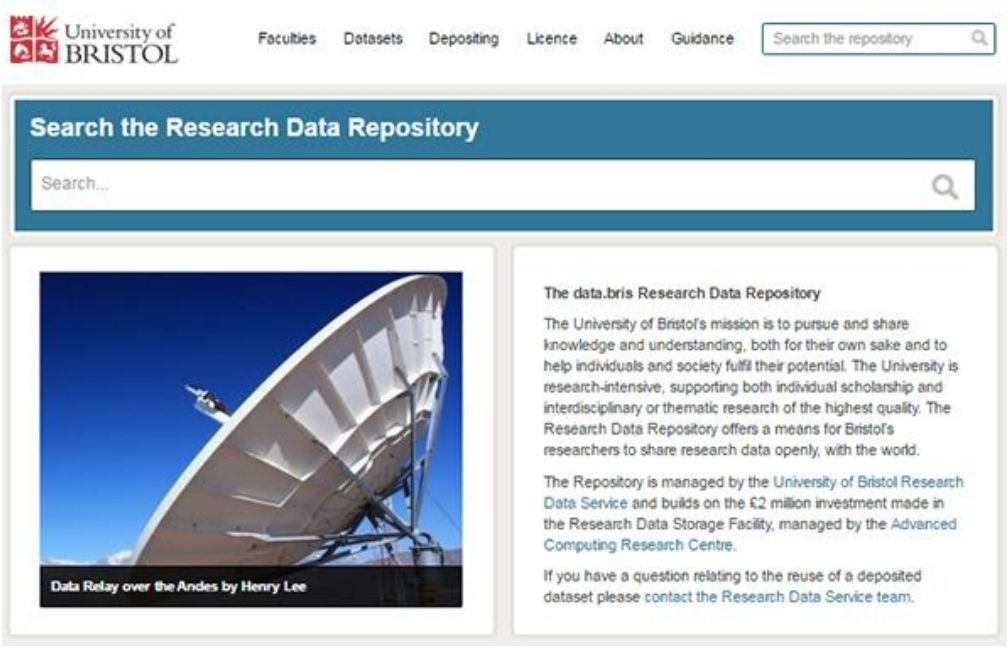

FIGURA 1. Repositório institucional
Fonte: University of Bristol http://data.bris.ac.uk/data/

Os repositórios institucionais ganharam notoriedade na década de 2000 com o surgimento de sistemas de software para sua implantação como Fedora e DSpace. Têm como objetivo recolher, administrar e manter a produção intelectual de uma instituição acadêmica ou de pesquisa e são plataformas projetados para a preservação e divulgação de publicações científicas (artigos, teses, documentos administrativos, etc.) gerados pelos membros da instituição. Facilitam a via verde para o Open Access (OA) ao proporcionar uma via que os pesquisadores autoarquivem todas as publicações, independentemente da fonte original que publicou o artigo.

Ao longo do tempo, os repositórios também permitiram o armazenamento de dados, facilitando a adição de descrições básicas e complexa de dados, e geralmente emitem identificadores que podem ser usados para citar e recuperar os dados. Alguns repositórios institucionais inclusive oferecem armazenamento ilimitado e, sendo apoiado por uma universidade, geralmente são administrados por bibliotecários.

Embora os repositórios institucionais ofereçam confiança, falta-lhes flexibilidade e controle. Muitos apresentam exigências rigorosas para aceitar o arquivo de dados de pesquisa usando formatos muito genéricos, faltam APIs falta a interoperabilidade com outros sistemas e muitos só utilizam um padrão muito geral de metadados como o Dublin Core e não suportam campos metadados de domínio de determinadas pesquisas ou tipo de dados específicos e vocabulários controlados.

4.2 Repositórios Temáticos

Os repositórios temáticos são aqueles que incluem dados de pesquisa de um campo disciplinar específico. Alguns repositórios temáticos bem sucedidos são arXiv, PubMed ou Eprints.

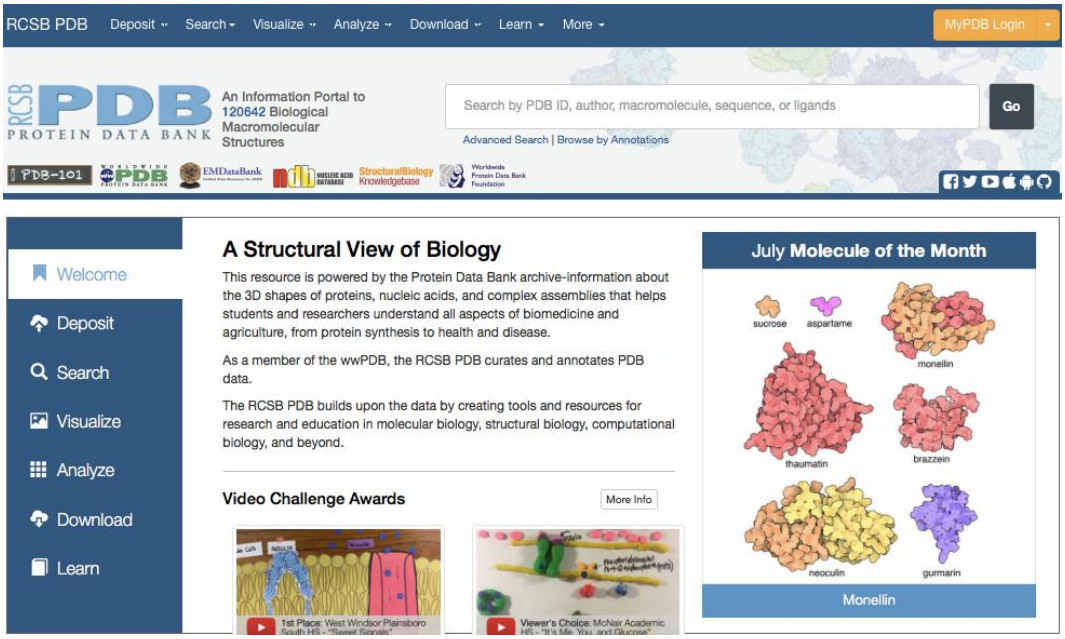

FIGURA 2. Repositorio Protein Data Bank
Fonte: Protein Data Bank www.wwpdb.org

Muitas disciplinas dispõem repositórios projetados especificamente para os tipos de dados em seu domínio. Alguns exemplos incluem o Protein Data Bank da Research Collaboratory for Structural Bioinformatics (RCSB), para formas em 3D de proteínas, ácidos nucleicos e conjuntos complexos; GenBank, para as sequências de DNA; o EMDataBank, com mapas 3D de microscopia eletrônica de densidade, modelos atômicos e metadados associados; os eCrystals para dados de cristalografia de raios-X, e o National Oceanographic Data Center (NODC) para dados oceanográficos.

Muitas vezes, estes repositórios temáticos apresentam ferramentas analíticas e de descobrimento disponíveis juntamente com os dados para fomentar sua reutilização. Alguns especialistas sugerem que os dados devem ser armazenados apenas em repositórios

temáticos, porque, dizem, permite o uso especializado de metadados e uma maior revisão e validação por especialistas na área. No entanto, nem todas as disciplinas dispõem repositórios de dados e a natureza específica e peculiaridade de muitos dados explica as dificuldades em encontrar o armazenamento em repositórios existentes.

4.3 Repositórios Editoriais

Os repositórios editoriais oferecem características semelhantes aos institucionais, mas com características especiais para comunidades específicas.

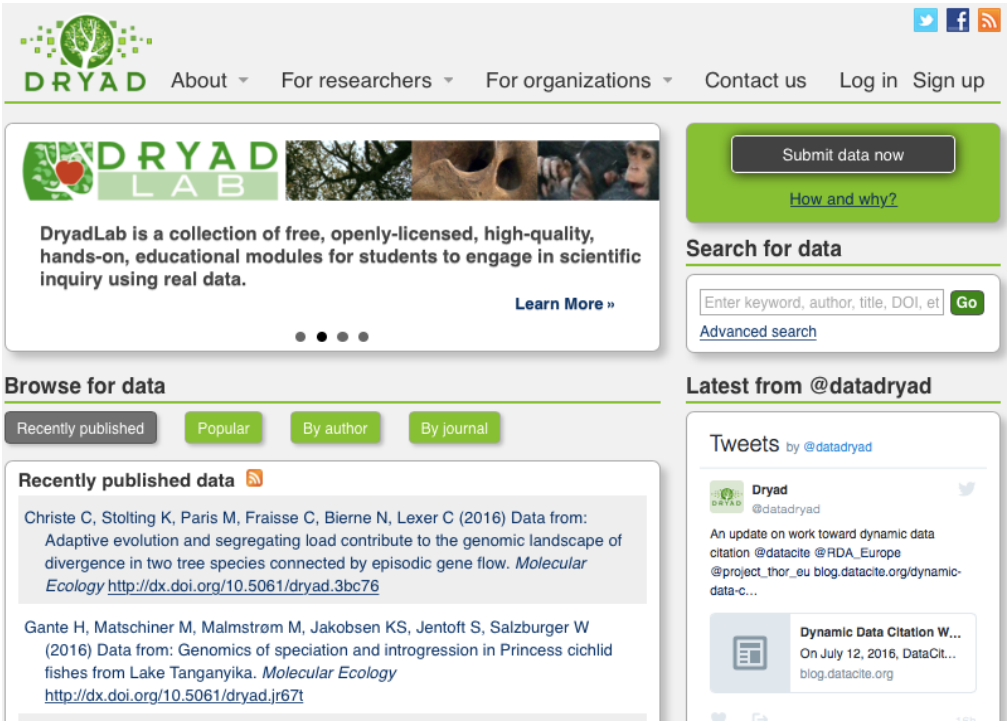

FIGURA 3. Repositorio Dryad (exemplo de registro)
Fonte: Dryad datadryad.org

Dryad é um repositório de dados de investigação científica e médica, de caráter internacional, procedente de periódicos científicos revisados por pares. Serve como repositório de várias disciplinas e também facilita o código DOI, atribuído pelo serviço EZID da Biblioteca Digital da Califórnia e registrado por DataCite. Dispõe de revistas tanto de acesso aberto como comerciais. Uma de suas principais características é a capacidade para acomodar qualquer tipo de dados órfãos. Além disso, Dryad minimiza a carga de apresentação de artigos, ou seja, o repositório faz uma leitura automática dos metadados fornecidos por revistas parceiras que fornecem informações bibliográficas para cada artigo antes da publicação.

DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333

4.4 Repositórios de Uso geral

Trata-se de repositórios que qualquer pesquisador pode usar, independentemente da sua filiação institucional, para preservar qualquer tipo de produção acadêmica. Os exemplos mais conhecidos são Figshare e Zenodo.

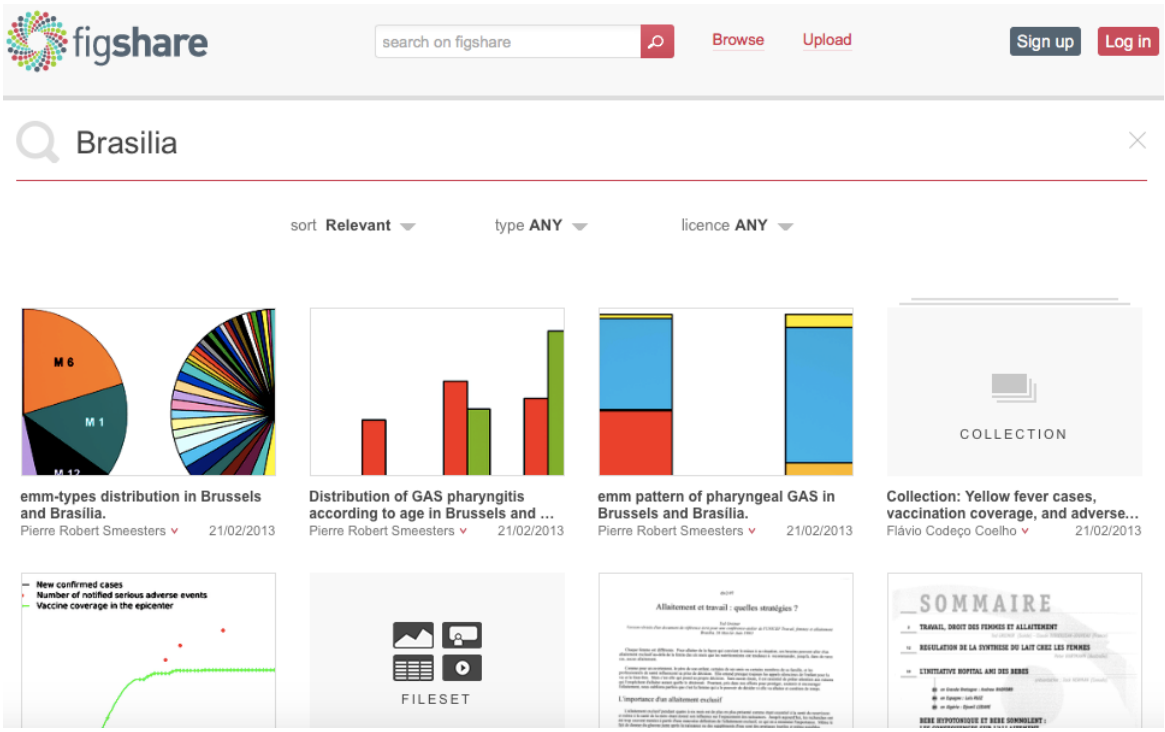

Figura 4. Figshare (página de resultados)
Fonte: Figshare https://figshare.com

Figshare é uma plataforma criada pela Digital Science para compartilhar e exibir os resultados de pesquisas multidisciplinares e é destinada a pesquisadores, cientistas, projetos e instituições. Atualmente está associada com o F1000 Research (um repositório de artigos científicos reconhecido pela comunidade acadêmica pelo seu alto nível de qualidade), colabora com PLOS (a maior revista de acesso aberto do mundo) e também com Plum Analytics (um serviço que quantifica o impacto das pesquisas publicadas). Todo o material publicado em Fisgshare é identificado com um DOI para facilitar a sua localização e a data. Na plataforma é possível localizar: apresentações, vídeos, cartazes, imagens, dados, artigos, etc. e a preservação de dados funciona com a tecnologia CLOCKSS, uma organização sem fins lucrativos que promove a parceria entre editoras e bibliotecas acadêmicas para arquivar de forma sustentável todos os conteúdos Web produzido no âmbito científico.

Os usuários podem integrar os dados do repositório com outros sites e blogs, copiando e colando um código. Os leitores podem realizar comentários sobre os conjuntos de dados e fazer download de arquivos para seus gestores de referência para uso posterior. O repositório também oferece a possibilidade de publicar resultados negativos ou experiências mal sucedidas para que outros pesquisadores economizem esforços sem ter que passar por testes já realizados e assim não desperdiçarem tempo de trabalho em determinados casos.

DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333

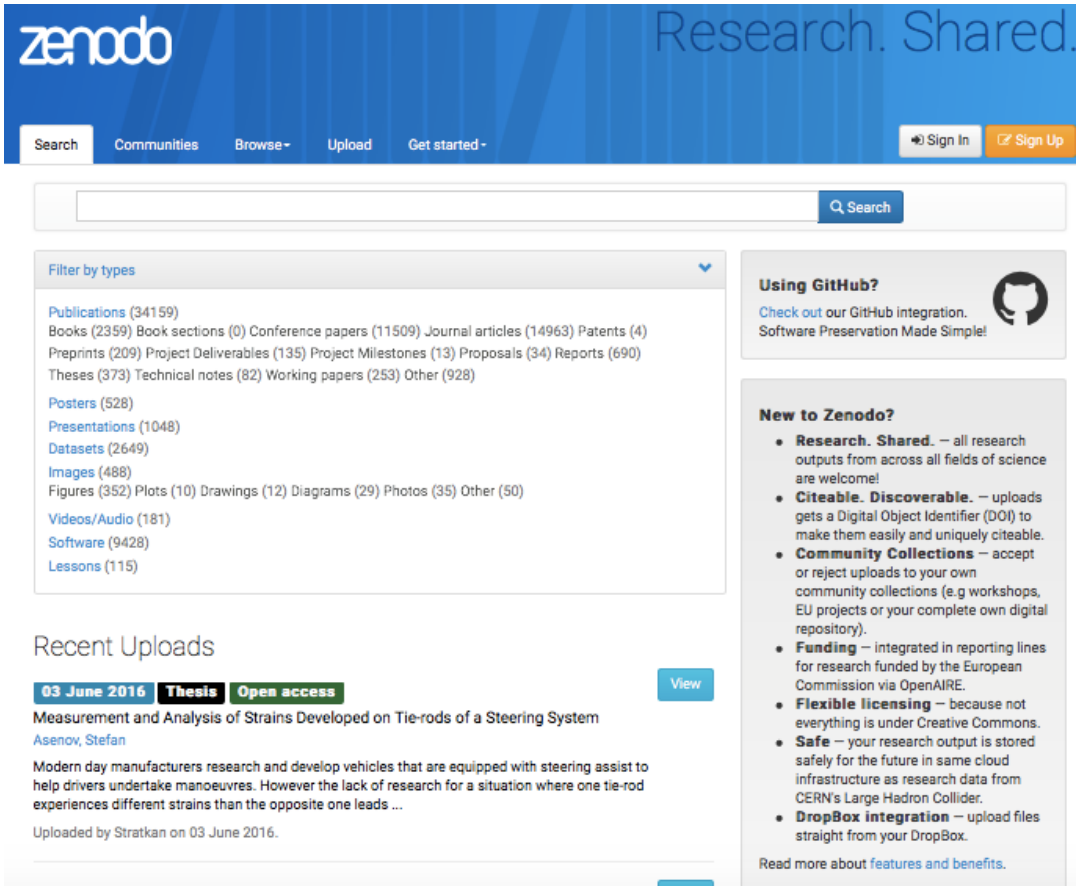

Figura 5. Repositorio Zenodo (página de inicio)
Fonte: Zenodo https://zenodo.org

Zenodo é uma iniciativa do portal OpenAIRE que dispõe infraestrutura adequada para acomodar conjuntos de dados e outros resultados de investigação de projetos europeus. Foi desenvolvido com a plataforma Invenio e desenvolvido pelo CERN, o centro que também é responsável por gerenciar a enorme quantidade de dados do Large Hadron Collider (LHC). Como no caso de Figshare, o acesso ao depósito é livre, aloca DOI e permite que conjuntos de dados disponível em BibTeX, EndNote e outros formatos bibliográficos. Os usuários podem adicionar metadados aos seus arquivos, muito mais detalhado do que em Figshare. Todos os dados são susceptíveis de serem recolhidos por outras plataformas através do protocolo OAI-PMH. Zenodo impulsiona o carregamento de dados através da comunicação com serviços como Mendeley, DropBox, CrossRef ou ORCID. Também inclui estratégias de preservação digital a longo prazo, permite estabelecer licenças flexíveis para gerenciar os direitos e permite aos usuários criar suas próprias coleções em um espaço próprio utilizando metadados sob licença CC0 dedicadas ao domínio público, sem restrições ou pedido de autorização, exceto para os endereços de e-mail. Além disso, sempre que permitido, outros usuários de Zenodo podem comentar seus arquivos, e uma característica interessante é que ele torna mais fácil para registrar com o identificador ORCID ou conta GitHub.

DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333

4.5 Repositorios próprios

Às vezes, os pesquisadores arquivam seus dados de pesquisa em um servidor pessoal ou de seus projetos, em nuvem. É possível encontrar diversas opções tecnológicas, incluindo a gratuita Dropbox, e outras comerciais, como Amazon Cloud Drive ou Microsoft Azure. Manter repositórios próprios depende da habilidade do pesquisador para acomodar as necessidades de seus projetos e levar a cabo acções adequadas de backup e replicação. Neste sentido, se trata de uma das opções menos recomendadas pela falta de garantias na organização, manutenção e preservação.

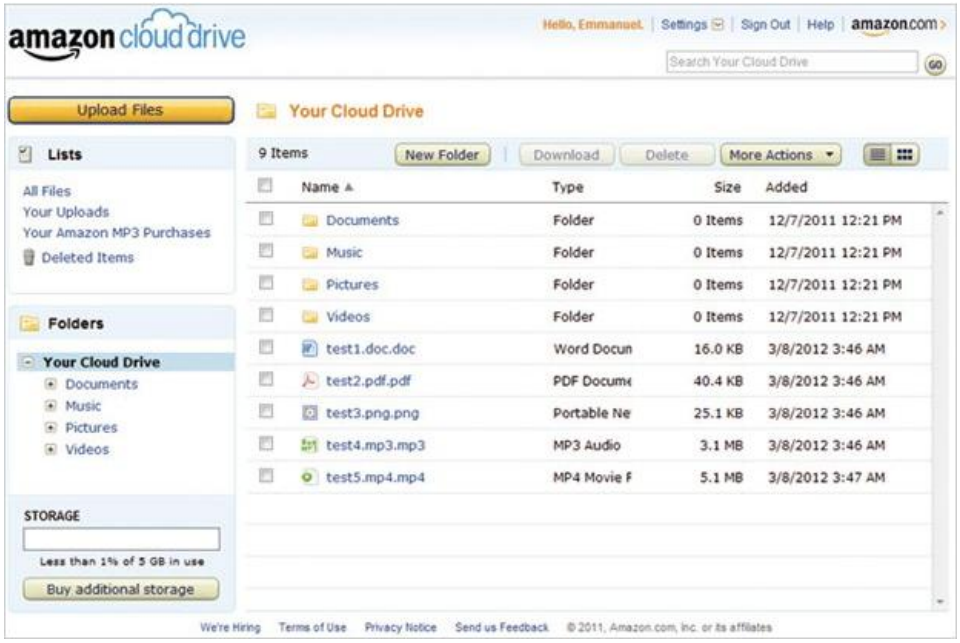

Figura 6. Repositorio Amazon Cloud Drive (conta de usuario)
Fonte: Amazon Cloud Drive

5 A REUTILIZAÇÃO DOS DADOS

Embora os requisitos para o compartilhamento de dados de pesquisa sejam relativamente novos, os sistemas de recolha e gestão ainda estão em desenvolvimento. Portanto, atualmente à pesquisa de um conjunto de dados em particular não é tão fácil quanto encontrar um artigo publicado. Para encontrar um conjunto de datos é recomendável começar a busca em artigos sobre o tema de interesse. Normalmente, os dados são depositados como material complementar ou link de um artigo.

Se a localização dos dados não se encontra no artigo, existem algumas alternativas. A primeira é a busca do currículo ou o site do autor para verificar se há qualquer referência à disponibilidade de dados em algum lugar. Se isso não funcionar também é possível entrar em contato com o autor para solicitar o acesso aos seus dados. As políticas de algumas revistas e agências de financiamento exigem uma cópia dos dados, desde que os dados não sejam

sigilosos. Nenhuma destas estratégias é infalível, uma vez que os dados mais antigos podem ser perdidos e os endereços de email mudam, mas pode ser uma boa estratégia para obter acesso aos dados que correspondem a um artigo.

Se a pesquisa é direcionada para os dados gerais de um tema e não para os dados de um artigo específico, a estratégia de busca deverá ser diferente. Um bom lugar para começar a procurar um tópico é um índice dos temas específicos de uma especialidade, sempre que existam. Por exemplo, o Integrated Ocean Observing System (IOOS) lista uma vasta gama de recursos marinhos na web e tem um motor de busca para ajudar a encontrar os dados específicos da sobre pesquisas dos oceanos. Esses índices não coletam dados necessariamente, mas apontam para uma série de recursos sobre um determinado tópico em conjunto com bancos de dados que também podem estar disponíveis em bibliotecas.

Na ausência de um banco de dados ou biblioteca, é possível considerar a busca em repositórios de dados que são populares em determinado campo particular e que podem ser localizados na lista re3data. Se deve ter em conta também as fontes externas dados, tais como agências governamentais, fundações de pesquisa, grupos de interesse especiais e outras organizações, já que muitas vezes fazem que os dados relacionados com suas atividades se tornem disponíveis. Por exemplo, a National Oceanic and Atmospheric Administration (NOAA) dos Estados Unidos é um excelente recurso para tudo o que se refere a dados relacionados com o clima. Tal como ocorre com qualquer tipo de informação, sempre é recomendável avaliar a fonte de dados para garantir sua credibilidade.

Finalmente, sabemos que, conforme o arquivamento público de dados se torne mais comum, a tendência é que fique mais fácil encontrar dados para serem reutilizados. O processo de investigação está em transição para um sistema de troca de dados, o que significa que muitos dos seus sistemas de negociação e reutilização estão em processo de desenvolvimento para que no futuro seja tão fácil encontrar os dados de um artigo como atualmente é para encontrar o próprio artigo.

6 DIREITOS DE REUTILIZAÇÃO DE DADOS

Uma vez que os pesquisadores encontram os dados que necessitam para suas pesquisas, devem considerar o que é permitido fazer com eles. Por exemplo, quando os dados são protegido por direitos autorais, a maioria das leis de copyright permitam a reutilização, com algumas exceções. Então, quando os dados da pesquisa são utilizados por outro pesquisador, é necessário analisar as condições de direitos de autorais e licenciamento. O tipo mais comum de reutilização de dados são os dados que têm licença aberta ou de domínio público. Os conjuntos de dados compartilhados publicamente frequentemente se encontram sob uma licença Creative Commons ou uma licença da *Open Data Commons*. A vantagem de usar os dados nestas condições é que a licença deve indicar claramente o que pode e não pode

fazer, tornando claros os direitos de reutilização. Tais licenças abertas, com exceções ocasionais, por exemplo a limitação da pesquisa comercial, permitem tanto a reutilização como voltar a publicar os dados da licença original. Basicamente, o uso de dados com uma licença aberta é a melhor opção, porque dispõe permissões relativamente extensas para fazer o que é necessário com os dados.

Os dados obtidos por meio de um financiamento também podem facilitar sua reutilização, embora muitas vezes com mais restrições para que o acesso seja aberto com uma licença. Como as agências financiadoras geralmente apresentam regras estipulando restrições de uso com os dados, se trata de seguir as disposições dos contratos. É necessário ter em mente que os contratos podem prevalecer sobre os direitos reconhecidos pelas leis de copyright. Por exemplo, se o pesquisador obtém um conjunto de dados em virtude de um contrato que proíbe a publicação de em acesso aberto, não será possível compartilhar.

Quando os dados não têm licenças que permitem sua reutilização, o uso se torna mais complicado. Neste caso, geralmente é permitido realizar pesquisas sobre o material, mas é necessário uma permissão para republicar os dados originais. Quando os dados são compostos de fatos naturais, geralmente é permitido publicar citando a fonte.

A documentação adequada é um dos requisitos mais importantes para reutilizar conjuntos de dados. Por exemplo, é difícil para o pesquisador de usar um conjunto de dados se é incapaz de determinar o significado dos nomes das variáveis. Por isso, é aconselhável começar buscando a documentação para a reutilização de um conjunto de dados. Os melhores conjuntos de dados são os que oferecem documentação detalhada, incluindo o arquivo README, ou um índice de conjuntos de dados. A melhor estratégia para superar uma documentação insuficiente é entrar em contato com o criador dos dados para mais informações.

Os erros em um conjunto de dados são outro problema que pode ser encontrado quando são utilizados dados de outro pesquisador. Tais erros incluem inconsistências, valores nulos, perdidos ou incorretos. Mesmo quando não forem encontrados erros em uma análise superficial dos dados, vale a pena realizar alguns testes básicos para verificar sua qualidade. Por exemplo, fazer um simples gráfico de dados é uma maneira fácil de verificar se há erros antes de usar um conjunto de dados para análise mais complexa. Outra vantagem de executar estes controles de qualidade é que se ganha uma melhor compreensão dos dados.

DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333 **ARTIGO**

7 REINICIO DO CICLO DE VIDA DOS DADOS

Reutilizar dados científicos é a última etapa do ciclo de vida dos dados, mas também o início de um novo ciclo. Os dados podem ser utilizados em novos projetos e ser incorporados as fase de coleta, análise e divulgação dos resultados. Os dados desempenham um papel importante em todos os processos e sua importância crescerá à medida que a partilha e reutilização se torne mais proeminente. Portanto, o pesquisador deve começar a pensar sobre seus dados como um produto que necessita ser preservado adequadamente, desde um projeto de pesquisa até sua finalização. Somente através de uma eficiente gestão de dados podemos perceber o potencial completo dos dados científicos.

Agora que chegamos ao fim do ciclo de vida dos dados e demonstramos o papel do bibliotecário neste processo, podemos nos detener um momento para refletir sobre o caminho percorrido ao longo do roteiro para o gerenciamento de dados de pesquisa. Contemplamos desde o planejamento para gerenciamento de dados, a documentação e a análise, até o compartilhamento e reutilização de dados, passando por diferentes tipos de armazenamento e segurança. Mediante a análise que fizemos, tanto pesquisador quanto bibliotecário podem encontrar ferramentas que ajudarão a gestionar adequadamente os dados científicos.

A medida que o pesquisador explore, investigue e incorpore estratégias de gestão de dados no fluxo de trabalho, recordará pontos importantes. Primeiramente, não é necessário fazer tudo de uma vez. A gestão de dados é a reunião de uma série de pequenos passos que se somam às boas práticas. É aconselhável trabalhar em uma prática por vez até que uma correta gestão de dados científicos se torne um hábito constante.

A gestão de dados também se torna mais fácil com o tempo. Isso ocorre em parte pela necessidade de fazer ajustes nos fluxos de trabalho de pesquisa adaptando novas estratégias. A tendência é que bons hábitos em relação aos dados se tornem parte da rotina da investigação. O objetivo é chegar a um ponto em que uma boa gestão de dados se converta simplesmente em um processo adicional de uma investigação.

Por último, é importante lembrar que o gerenciamento de dados é um processo vivo. Não é uma meta que se deve chegar para que o pesquisador não tenha que voltar a fazer a gestão dos dados. Um gerenciamento de dados correto requer esforço, mas esse esforço será recompensado mais tarde, quando o pesquisador não tenha que passar horas tentando encontrar, entender ou reutilizar os seus dados. Ao tornar a gestão de dados consciente e contínua, o pesquisador pode reduzir significativamente as frustrações diárias que surgem para quem trabalha com dados de pesquisa. Os dados devem trabalhar para o pesquisador e não contra ele; a pesquisa científica já por si mesma suficientemente difícil.

DOI: http://dx.doi.org/10.20396/rdbci.v14i3.8646333

CONSIDERAÇÕES FINAIS

O planejamento do ciclo de vida dos dados é uma ferramenta importante para começar a delinear a infraestrutura que o pesquisador deseja oferecer. Cada etapa do ciclo de vida requer diferentes níveis de organização e este modelo pode ser uma maneira útil de pensar sobre o que é possível e o que se quer fazer.

Se antes os pesquisadores produziam conhecimento buscando documentos em diferentes repositórios, hoje os dados científicos ganharam papel de destaque para o avanço da produção científica. Como consequência, esta nova importância atribuída aos dados científicos adquiriu novas demandas para preservar e recuperar os dados científicos. Conforme demonstramos, o bibliotecario possui o perfil necessário para auxiliar pesquisadores no desenvolvimento de um projeto de gestão de dados e indicar as ferramentas necessárias para este procedimento. Além disso, é imprescindível que ocorra a seleção adequada de um repositório conforme os requisitos de determinados conjuntos de dados, processo no qual ambos, bibliotecarios e pesquisadores podem trabalhar juntos.

REFERÊNCIAS

CARLSON, Jake R. Demystifying the data interview: Developing a foundation for reference librarians to talk with researchers about their data. *Reference Services Review* v. 40, n. 1. 2012. p. 7-23. Disponível em: < http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1186&context=lib_research >. Acesso em: 03 ago. 2016.

European Commission. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. v. 1.0. Dec. 2013.

GARRITANO, Jeremy R.; CARLSON, Jake R. A Subject librarian's guide to collaborating on e-science projects. *Issues in Science and Technology Librarianship*, n. 57, 2009.

LAGE, Kathryn; LOSOFF, Barbara; MANESS, Jack. Receptivity to library involvement in scientific data curation: a case study at the University of Colorado Boulder. *Libraries and the Academy*, v. 11, n. 4. 2011. p. 915–937.

SCIENTIFIC DATA. Disponivel em: < http://nature.com/scientificdata >. Acesso em: 03 ago. 2016.

3TU. Datacentrum, 2015. 3TU. Datacentrum. Disponivel em: < http://datacentrum.3tu.nl/. en/home. Acesso em: 03 ago. 2016.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil

ISSN 1678-765X