DOI: 10.5433/1981-8920.2016v21n2p90

ALGUMAS CONSIDERAÇÕES SOBRE OS REPOSITÓRIOS DIGITAIS DE DADOS DE PESQUISA

SOME CONSIDERATIONS ON THE DIGITAL REPOSITORIES OF RESEARCH DATA

Luis Fernando Sayão*
Luana Farias Sales**

RESUMO
Introdução: A pesquisa científica contemporânea no seu compromisso por buscar novos conhecimentos e novas descobertas produz e utiliza intensivamente dados digitais de pesquisa. Nesse contexto de mudanças, os dados deixam de ser simples subprodutos das atividades de pesquisa e se tornam recursos informacionais de primeira grandeza, caracterizando um novo paradigma científico pautado pelo compartilhamento, amplo acesso e reuso de dados.
Objetivo: Identificar o papel dos repositórios digitais de dados nos novos cenários de pesquisa científica e apresentar um panorama das suas principais características, categorias, benefícios, funções e infraestruturas.
Metodologia: Analisa a literatura da área e os principais sistemas que dão sustentação a infraestruturas de acesso e gestão de dados de pesquisa.
Resultados: O presente ensaio demonstra que para os dados e coleções de dados de pesquisa transmitam conhecimento no tempo e no espaço e sejam reusados é necessária a implantação de uma infraestrutura tecnológica e gerencial, permanente e sustentável, que permita que eles sejam cuidados ao longo de todo o seu ciclo de vida. No centro dessa infraestrutura estão os repositórios digitais de dados de pesquisa, que são sistemas voltados para apoiar a seleção, catalogação, arquivamento, acesso e compartilhamento de dados de pesquisa.
Conclusão: Pela sua importância como recurso informacional, os repositórios de dados se tornam rapidamente parte essencial das infraestruturas de pesquisa em escala global, tornando visível e aberta para toda a sociedade uma parcela importante da atividade de pesquisa. Nessa direção, se tornam um desafio relevante para a Ciência da Informação e para a Biblioteconomia.

Palavras-chave: Dados de pesquisa. Repositório digital de dados de pesquisa. Gestão de dados de pesquisa. Curadoria digital. Ciência aberta.

*Doutor em Ciência da Informação pela Universidade Federal do Rio de Janeiro (UFRJ/IBICT). Centro de Informações Nucleares (CNEN/CIN). E-mail: lsayao@cnen.gov.br.
**Doutora em Ciência da Informação pela Universidade Federal do Rio de Janeiro (UFRJ/IBICT). Pesquisadora do Instituto de Engenharia Nuclear (CNEN/IEN). E-mail: lsales@ien.gov.br.

1 INTRODUÇÃO

No ciclo de geração de conhecimento científico há uma parcela considerável do trabalho de pesquisa que necessita de infraestruturas informacionais formalizadas para se tornarem visíveis para as próprias comunidades acadêmicas, para as instituições de pesquisa e agências de fomento e para a sociedade como um todo. Trata-se dos dados digitais de pesquisa que muito rapidamente deixam de ser meros subprodutos das atividades de pesquisa e se tornam um foco de grande interesse para todo o mundo científico.

Nesse território da pesquisa contemporânea, reordenado pela geração e uso intensivo de dados, os meios de compartilhamento e o amplo acesso aos dados de pesquisa criam pontos de inflexões nas metodologias de identificação de novos fenômenos, na validação e na reprodutibilidade das pesquisas e nas formas de socialização dos pesquisadores. Este cenário de grandes novidades abre perspectivas inéditas para descobertas em todas as áreas do conhecimento, que vai da Astrofísica a Linguística, delineando um novo paradigma científico.

Porém, de forma diferente das publicações acadêmicas que falam por si próprias e têm explícitos os seus conteúdos, as coleções de dados de pesquisa para revelarem e transmitirem conhecimento no tempo e no espaço, e a partir daí serem interpretadas, sintetizadas e reanalisadas em contextos diversos – e diferentes para os quais foram geradas e coletadas originalmente - precisam de ações específicas que permeiam todo o seu longo ciclo de vida. Essas ações se iniciam no planejamento da criação dos dados, passam pela organização desses dados em coleções identificadas por meios de referências estáveis e padronizadas e vão até o arquivamento de longo prazo para os dados reconhecidos como de valor permanente ou que são referencias universais.

Nessa direção, para lidar com os seus desafios e enigmas, as ciências orientadas por dados necessitam não somente de infraestruturas físicas – como aceleradores de partículas, navios de pesquisa e computação em grade. Elas devem dispor também de sistemas de informação que incorporem arquivamento e compartilhamento de coleções de dados (PANPEL et al.,

2013), além de uma gestão dinâmica que permita adicionar valor a esses dados tendo como perspectiva a usabilidade, a identificação de novas relações e padrões e a pesquisa interdisciplinar. Idealmente, esses sistemas digitais de informação – que reúnem dados e publicações acadêmicas - devem estar imbricados aos demais recursos infraestruturais voltados para a pesquisa, compondo o que pode ser chamado de ciberinfraestrutura de pesquisa, "uma nova forma de cultura científica que se sustenta em uma robusta infraestrutura tecnológica de alto nível" (PÉREZ-GONZÁLES, 2010, p. 3).

As ações que coletivamente permeiam o ciclo de vida dos dados de pesquisa vêm sendo chamadas de gestão de dados de pesquisa, e se consolidam, de maneira ideal, em consonância com padrões de ampla aceitação, condicionantes de domínios disciplinares, requisitos estabelecidos pelos pesquisadores e, quando existentes, políticas institucionais e diretrizes de alcance nacional e internacional.

Para apoiar a execução dos processos de gestão é necessário um arcabouço tecnológico e gerencial que compreenda todo o ciclo de vida dos dados. No centro desse arcabouço estão os **repositórios digitais de dados de pesquisa** que, por muitas razões e demandas, rapidamente se tornam parte essencial da infraestrutura mundial de pesquisa. Duas dessas demandas são determinantes para a ampliação das ações em torno da gestão de dados e para o seu ordenamento: as políticas mandatórias das agencias financiadoras de pesquisa e a incorporação pelas instituições de pesquisa e pela sociedade em geral dos valores e princípios preconizados pela Ciência Aberta.

Diante do pressuposto de que dados de pesquisa são, em grande parte, resultados de pesquisas financiadas com dinheiro público, as instâncias governamentais e as agências de fomento começam a definir regras compulsórias para a gestão e para a disponibilização desses dados para acesso aberto e contínuo; convergindo para a mesma direção, o movimento da Ciência Aberta – que considera o conhecimento científico como um patrimônio da humanidade - amplia as exigências em torno da questão de dados abertos, incluindo metodologias, ferramentas, modelos, *softwares* e tudo mais que garanta os princípios de transparência, reprodutibilidade e autocorreção da ciência (KON, 2013; THE ROYAL SOCIETY, 2012).

Em face da amplitude das exigências dos pesquisadores, das agências de fomento e da sociedade, expressas pelos pressupostos da Ciência Aberta, as instituições de pesquisa começam a expandir os seus serviços de repositório digital – que até então estavam voltados preferencialmente para publicações digitais (*eprints*) - para o domínio heterogêneo e complexo dos dados de pesquisa; como desdobramento desse fato, começam a planejar e implantar sistemas de repositório de dados de pesquisa nas mais diversas configurações, plataformas tecnológicas e modelos de gestão.

Nesse contexto de transição, dois problemas se colocam: por um lado os pesquisadores necessitam de infraestruturas que assegurem o máximo de confiabilidade, estabilidade e acessibilidade e que facilitem o trabalho de arquivamento, compartilhamento e reconhecimento de autoria para os seus dados; por outro lado, esses mesmos pesquisadores precisam encontrar coleções de dados de pesquisa, saber como acessá-las e sob que condições podem reutilizar esses dados e assim dar prosseguimento às suas pesquisas confiando na autenticidade e proveniência dos dados coletados ou gerados por outros pesquisadores.

Tomando como ponto de partida as questões acima como pretexto para uma reflexão sobre o papel dos repositórios nos atuais cenários de pesquisa científica, o presente ensaio analisa parte da literatura da área, e apresenta, como resultado, uma sistematização dos principais categorias e características dos repositórios de dados, alinhando seus benefícios e funções; são considerados também alguns sistemas que estão à volta desses repositórios.

2 DEFINIÇÕES E CONTEXTOS

O termo “dado de pesquisa” tem uma amplitude de significados que vão se transformando de acordo com domínios científicos específicos, objetos de pesquisas, metodologias de geração e coleta de dados e muitas outras variáveis. Pode ser o resultado de um experimento realizado num ambiente controlado de laboratório, um estudo empírico na área de ciências sociais ou a observação de um fenômeno cultural ou da erupção de um vulcão num determinado momento e lugar. Dados digitais de pesquisa ocorrem na forma de

diferentes tipos de dados, como números, figuras, vídeos, *softwares*; com diferentes níveis de agregação e de processamento, como dados crus ou primários, dados intermediários e dados processados e integrados; e em diferentes formatos de arquivos. Essa diversidade vai sendo delineada pelas especificidades de cada disciplina, suas condicionantes metodológicas, protocolos e seus objetivos, e se torna um desafio, mesmo para o pesquisador, pelo alto grau de contextualização necessário, definir precisamente o que é dado de pesquisa de uma forma transversal aos diversos domínios disciplinares (BORGMAN, 2010; PAMPEL et al., 2013).

O relatório da *National Science Board* (2005) categoriza os dados de pesquisa em: **dados experimentais**, resultados de estudos em ambientes controlados de laboratórios; **dados computacionais**, que são produtos da execução de modelos computacionais que simulam uma dada realidade; e **dados observacionais** que são resultados de observações de fenômenos que se desenrolam em lugares e tempos específicos. Mais que uma definição essa classificação – já considerada clássica - possibilita uma compreensão mais clara e contextualizada do que pode ser dado de pesquisa. Além do mais, essa categorização ajuda a definir o grau de gestão e de curadoria necessários para manter a usabilidade de cada tipo de dados, que é algo relevante para o presente estudo. Por exemplo, os dados observacionais precisam ser preservados para sempre, pois não há como coletá-los novamente, e, portanto, precisam de uma gestão de longo prazo mais intensa e cuidadosa; por outro lado, no âmbito da simulação os modelos computacionais é que precisam ser preservados e não os dados resultantes da execução de códigos gerados a partir desses modelos.

Dependendo do ponto de observação, quase tudo que é gerado e coletado no ambiente de pesquisa pode ser considerado dado de pesquisa. Esse fato coloca a importância de examinar o conceito de dados de pesquisa ancorado por um sistema de informação que pode formalizá-lo como objeto de informação por meios de fluxos de trabalho apropriados e processos de representação e ressignificação que os tornem passíveis de servirem a diferentes disciplinas ou comunidades científicas, segundo os respectivos conceitos de dados de pesquisa que cada disciplina e comunidade apropriam.

Isto fica mais evidente sob a perspectiva de interpretabilidade e usabilidade dos dados. “Dados de pesquisa não tem valor sem seus metadados e documentação apropriada que descrevem seus contextos e as ferramentas usadas para criá-los, armazená-los, adaptá-los e analisá-los” (KINDLING; SCHIRMBACHER, 2013).

Dessa forma, dados digitais de pesquisa são objetos digitais que precisam de informação de representação (identificador, metadados, proveniência, documentação, caderno de laboratório, etc.) para se tornarem objetos informacionais (CCSDS, 2012). O que significa dizer que disponibilizar os dados na web, sem a necessária contextualização estrutural e semântica, impossibilita a interpretação e o reuso desses ativos; impossibilita a transmissão do conhecimento que ele porta, a reinterpretação em outros contextos e, consequentemente, reduz o seu valor para a pesquisa interdisciplinar. Os elementos de representação são atribuídos pelos processos de gestão de dados de pesquisa, que “resumem todas as ações necessárias para tornar os dados passiveis de serem descobertos, acessíveis e compreensíveis ao longo do tempo: organização, documentação, armazenamento, compartilhamento e arquivamento” (LEIDEN UNIVERSITY, 2015, p. 1). A gestão tem como infraestrutura central, conforme já afirmado, o repositório digital de dados de pesquisa.

Além de oferecer uma base tecnológica para a execução dos processos de contextualização dos dados, os repositórios têm um papel importante nas interações que envolvem a validação do trabalho de pesquisa e na própria dinâmica social da comunicação científica. A possibilidade de se ter os dados de pesquisa disponíveis *on-line*, indexados, documentados e anotados relativos a uma pesquisa publicada ou pré-publicada num artigo acadêmico, redimensiona a revisão por pares, estendendo-a a uma comunidade mais ampla e conectada em rede. “Um repositório permite exame, prova, revisão, transparência de resultados de pesquisa por outros especialistas que vão além da revisão por pares do artigo acadêmico publicado” (UZWYSHYN, 2016, p. 1).

Há ainda uma perspectiva documental que escapa aos repositórios de *eprints* e bibliotecas tradicionais e digitais. A pesquisa científica – por sua orientação para a descoberta – confere maior visibilidade aos produtos finais

do seu ciclo de vida: publicações, inventos, patentes, protótipos. Entretanto, a ciência trilha uma trajetória de erros e acertos que somados impulsionam o seu progresso. Essa trajetória aparentemente errática pode não estar registrada nas publicações, porém tem um papel relevante na contextualização do fazer científico. As coleções de dados coletados e gerados pelos pesquisadores bem como as suas versões e linhagens contam melhor essa história e podem ser arquivadas naturalmente nos repositórios que possuem modelos de dados adequados para tal; além do mais, os repositórios de dados permitem a publicação de dados negativos que anteriormente estavam ocultos e sem um ambiente apropriado para serem arquivados e disseminados. Esses dados correspondem a experimentos que não deram certo. Isto permite que outros pesquisadores evitem trilhar vias sem saída, que pesquisadores anteriores tentaram, e achem seu caminho na direção de territórios mais férteis (UZWYSHYN, 2016).

Nessa perspectiva, os repositórios de dados de pesquisa são infraestruturas de base de dados desenvolvidas para apoiar todo o ciclo da gestão de dados de pesquisa, incluindo as ações mais dinâmicas e contundentes sobre os dados, que coletivamente são chamadas de curadoria de dados de pesquisa, que visam adicionar valor aos dados, avaliando, formatando, agregando e derivando novos dados. Para tal, os repositórios desempenham inúmeras funções que são resumidas na seção que se segue.

3 BENEFÍCIOS E FUNÇÕES DO REPOSITÓRIO DE DADOS

Os repositórios de dados de pesquisa têm como objetivo fundacional garantir o acesso contínuo e aberto - agora e no futuro - aos resultados de pesquisa que se manifestam na forma de dados, e que são considerados parte importante do patrimônio digital da humanidade, conforme enfatiza a página web do SURF Foundation[1]. As ações dos repositórios de dados nessa direção trazem benefícios relevantes para o mundo da pesquisa científica e para toda a

[1]<https://www.surf.nl/en/themes/research/research-information/research-repositories/index.html>

sociedade. Partindo da analise preliminar de Sales (2014), a seguir são apresentados alguns dos benefícios mais perceptíveis:

- VISIBILIDADE DOS DADOS
  Amplia a visibilidade dos dados de pesquisa permitindo que eles sejam consultados e citados mais frequentemente (geralmente só é disseminada pelos canais formais a fração de dados – cerca de 10% - que está registrada nos artigos publicados).
- COMPARTILHAMENTO DE DADOS
  Os repositórios pela sua capacidade de agregação e organização de recursos informacionais dispersos no tempo e no espaço, e como instrumento de socialização de comunidades e grupos pesquisadores ao redor desses recursos, tornam-se um dispositivo importante de troca de experiências e compartilhamento de dados.
- CRÉDITO AO AUTOR DOS DADOS
  Os repositórios de dados tornam possível identificar as coleções de dados e seus autores de forma unívoca e persistente, permitindo que os autores sejam reconhecidos, citados, avaliados e recompensados pelo trabalho intelectual de coleta, geração e organização dos dados.
- PRESERVAÇÃO DIGITAL
  Oferece um ambiente tecnológico, gerencial e de padronização propício para a preservação de longo prazo dos dados de pesquisa de valor contínuo, especialmente para os dados observacionais.
- MEMÓRIA CIENTÍFICA E TRANSPARÊNCIA
  Contribui para a formação da memória científica das instituições no que diz respeito aos dados, complementando os repositórios institucionais que estão focados nas publicações acadêmicas; na qualidade de registro das atividades de pesquisa das instituições, contribui também com os princípios de transparência, tão em voga nos tempos atuais.

- SEGURANÇA DOS DADOS
  Oferece sistema de armazenamento seguro, esquemas de *backup* e segurança física que se contrapõem ao armazenamento informal em mídias portáteis e computadores pessoais frequentemente usados pelos pesquisadores.
- DISPONIBILIDADE
  Permite que os dados estejam disponíveis *on-line* para serem acessados, baixados, visualizados e processados por pessoas ou por sistemas.
- CURADORIA DIGITAL
  Proporciona um ambiente apropriado para os processos de avaliação, de adição de valor, reformatação, agregação e recriação de dados promovidos pela curadoria digital.
- SERVIÇOS INOVADORES
  Abre possibilidades de criação de novos serviços de informação para pesquisadores, gestores e financiadores de pesquisa a partir da análise e integração dos dados arquivados com fontes internas e externas à instituição.
- REUSO DOS DADOS
  Aumenta o grau de reuso e reinterpretação dos dados possibilitando a realização de novas pesquisas de caráter interdisciplinar; minimiza a duplicação de esforços e otimiza os investimentos na coleta e geração de dados.
- REDES DE REPOSITÓRIOS
  Permite por meio de protocolos de interoperabilidade, como o OAI-PMH, a formação de redes de repositório dados; abre a possibilidade de inserção dos repositórios de dados às redes interoperáveis definidas pelo padrão *Linked Data*.
- INDICADOR DE QUALIDADE E PRODUTIVIDADE DA INSTITUÇÃO
  As coleções de dados organizadas e arquivados no repositório são evidências da qualidade e da relevância das atividades de pesquisa da instituição, atestando a sua produtividade e seu valor acadêmico.

O repositório de dados, na qualidade de um sistema digital que integra diversas funções, tem como perspectiva oferecer um ambiente dinâmico e flexível – principalmente pela natureza heterogênea dos dados - para dar apoio à execução dos processos de gestão de dados de pesquisa. Grande parte das funções que se desenrolam nesse ambiente se enquadra mais no escopo administrativo e biblioteconômico do que no escopo tecnológico.

O elenco de funções pode ser bastante diversificado, mas de forma ideal devem estar orientadas por uma política institucional explicita de gestão de dados de pesquisa, registrada em um documento disponível publicamente que deve refletir as opções e condicionantes das instituições, as demandas dos grupos de pesquisadores e as singularidades dos domínios disciplinares onde atua.

A figura 1 representa um fluxo de gestão de dados de pesquisa que se realiza num repositório genérico, que opera de acordo com uma política institucional de gestão de dados.

Figura 1 – Fluxo de gestão de dados de pesquisa

Fonte: Autores

O quadro 1 detalha as possíveis funções desempenhadas em cada etapa considerada. Fica claro que em comparação com um repositório de *eprints* as funções, as descrições, os padrões e os controles são mais numerosos e complexos, no entanto, essa complexidade varia de acordo com os ambientes disciplinares considerados e com a politica adotada pela instituição (SAYÃO, SALES, 2015).

Quadro 1 – Funções de um sistema de gestão de dados de pesquisa.

CAPTURA DE DADOS Inserção de coleções de dados – primários ou derivados - provenientes de experimentos, simulações, observações, questionários, levantamentos, etc. Os dados podem ser incorporados ao repositório pelos próprios autores - por autossubmissão - ou por equipes especializadas vinculadas ao serviço.	
FUNÇÕES	• Seleção dos dados passíveis de serem arquivados; • Verificação do enquadramento no escopo do repositório; • Verificação dos formatos de arquivos aceitáveis para submissão; • Verificação dos direitos associados às coleções (copyright e licenças); • Verificação de dados sensíveis (dados não anonimizados, confidenciais, pessoais); • Verificação do volume e quantidade de arquivos; • Verificação dos metadados gerais e disciplinares que acompanham os dados; • Normalização para elenco de formatos padronizados aceitos para arquivamento e disseminação; • Controle de qualidade dos dados; • Definição de tempo de embargo
CATALOGAÇÃO DAS COLEÇÕES DE DADOS Descrição, atribuição de metadados e inclusão de documentação que assegurem que os dados possam ser acessados e interpretados no tempo e no espaço.	
FUNÇÕES	• Atribuição de: metadados descritivos, estruturais, administrativos, técnicos (que inclui os relativos às dependências técnicas dos objetos digitais) • Atribuição de metadados de preservação, que assegurem a proveniência, autenticidade e integridade dos dados ao longo do tempo; • Uso de taxonomias especializadas e disciplinares; • Atribuição de identificador persistente (DOI, Handles, UNF, URN, etc.) que permita que os dados possam ser localizados de forma persistente e citados como as publicações acadêmicas; • Identificação do autor (ORCID ID, Scopus Author ID, ResearcherID, etc.) • Inclusão de documentação sobre os dados, incluindo descrição do projeto, dos arquivos e dos parâmetros; cadernos de laboratório e de campo, protocolos de pesquisa ou metodologia, etc.; • Vinculação (por links) a publicações e a dados relacionados internos e externos ao repositório
ARQUIVAMENO E PRESERVAÇÃO Arquivamento seguro que garanta a gestão de curto e longo prazo das coleções de dados orientados por um plano/politica de preservação digital	
FUNÇÕES	• Armazenamento em sistemas seguros; • Gestão da preservação de curto prazo (backups, backups redundantes offsite; checagem de integridade, armazenamento seguro, criptografia, compressão); • Gestão de longo prazo (migração, emulação, reformatação para formatos padronizados, aplicação de normas pertinentes (OAIS, TRAC), informação de fixidade voltada para validar a autenticidade e integridade de um objeto digital (checksums, assinatura digital) • Implementação de trilhas de auditoria.
INTEROPERABILIDADE Intercâmbio e compartilhamento e linkage com outros repositórios de dados e outros sistemas de informação (repositórios institucionais, bibliotecas digitais de publicações acadêmicas, editoras científicas)	
FUNÇÕES	• Disponibilização de metadados segundo o protocolo OAI-PMH; • Agregação para formação de publicações ampliadas segundo o padrão OAI-ORE; • Uso dos padrões, web service, linked data e outros

	• Empacotamento de metadados para intercâmbio segundo o padrão METS;
RECUPERAÇÃO, ACESSO E REUSO Interface web para a descoberta, acesso e downloading de coleções de dados relevantes para o usuário ou para aplicações computacionais, como visualização e mapeamento, que podem prover serviços a partir dessas coleções; vinculado a uma política de acesso estabelecida pela instituição que inclui: tempo de embargo, direito de acesso, pagamentos, restrições sobre determinadas coleções, acesso somente aos metadados; registros de usuários e termos de uso dos dados.	
FUNÇÕES	• Disponibilização de interfaces web para recuperação, acesso e downloading; • Oferta de aplicações e serviços sobre as coleções;

Fonte: Autores

4 TIPOS DE REPOSITÓRIOS

Há um grau de consenso entre os autores que se dedicam à análise dos repositórios de dados de pesquisa que eles podem ser caracterizados – com algumas sobreposições e diferenças - como institucionais, disciplinares, multidisciplinares e orientados por projetos. Baseado no esquema enunciado por Pampel e seus colaboradores (2013), analisa-se a seguir as possíveis diferenças e limites entre as categorias de repositórios de dados.

Repositórios institucionais de dados de pesquisa

Essa categoria de repositórios de dados é caracterizada por ser gerenciada e funcionar no âmbito de uma instituição acadêmica, como universidades ou institutos de pesquisa, e são voltados para arquivar dados que são, geralmente, provenientes unicamente das atividades acadêmicas dessas instituições. Algumas vezes são os próprios repositórios institucionais que estendem seus modelos de dados para incluir dados de pesquisa, como é, por exemplo, o caso do Carpe Dien[2] do Instituto de Engenharia Nuclear (IEN/CNEN). Porém, na maioria dos casos, os repositórios de dados são plataformas independentes - especialmente pela necessidade de esquemas mais ricos de metadados -, mas que podem estabelecer *links* entre os seus recursos e os recursos dos repositórios de *eprints*, estabelecendo novas formulações de publicações acadêmicas (SALES, 2014).

[2] http://carpedien.ien.gov.br/

Considerando o amplo escopo das pesquisas desenvolvidas nas universidades e em alguns institutos de pesquisa, esses repositórios são também **multidisciplinares**, na medida em que armazenam dados provenientes das diversas vertentes de pesquisa da instituição. Na qualidade de repositório institucional, cumprem também o papel de registrar a parte da **memória acadêmica** da instituição circunscrita pela geração de dados. O Edinburgh Data Share (UK)[3] é um exemplo de repositório institucional e multidisciplinar voltado para os dados de pesquisa gerados pelas pesquisas da Universidade de Edinburgh.

Repositórios disciplinares de dados de pesquisa

São repositórios voltados para o arquivamento de domínios específicos de pesquisa como física de partículas ou ciências ambientais. Em algumas condições se orientam para tipos particulares de dados, como por exemplo, a BioModels Database[4] que é um repositório voltado para arquivamento, descoberta e intercâmbio de modelos computacionais na área de biologia. Estes modelos – essenciais para o desenvolvimento de *software* de simulação - são descritos em periódicos científicos revisados por pares, e enfatizam o fato de que os dados gerados por simulação, ou seja, por *software*, não são diferentes dos dados gerados por outras formas de pesquisa (THE ROYAL SOCIETY, 2012).

Outros exemplos marcantes, de alcance planetário e que caracterizam bem a categoria, é o GenBank[5] e o PANGAEA[6]: o GenBank é uma base de dados de sequenciamento genético que suporta anotação bibliográfica e biológica que incorpora todas as sequencias de DNA publicamente disponíveis (BENSON, et al., 2013); por sua vez o PANGAEA – Data Publisher for Earth & Environmental Science - é uma biblioteca digital aberta voltada para o armazenamento, publicação e distribuição de dados georeferenciados

[3] http://datashare.is.ed.ac.uk
[4] https://www.ebi.ac.uk/biomodels-main/
[5] http://www.ncbi.nlm.nih.gov/genbank/
[6] https://www.pangaea.de/

provenientes das pesquisas no campo da Ciência do Sistema Terrestre. A biblioteca armazena mais de meio milhão de coleções de dados originados em todos os domínios das Geociências.

O que fica patente é que os repositórios disciplinares ou temáticos, como são denominados por alguns autores, são extremamente variados e heterogêneos, refletindo a multiplicidade de disciplinas e a diversidades de dados gerados no contexto da pesquisa científica mundial.

Repositórios multidisciplinares de dados de pesquisa

São repositórios que reúnem coleções de dados coletados ou gerados por atividades de pesquisa em várias áreas de conhecimento. Conforme já observado, uma grande parcela dos repositórios institucionais vinculados às universidades – pela natureza multidisciplinar dessas instituições – recai nessa categoria também. Para a finalidade do presente estudo, estão qualificados como primariamente multidisciplinares, os repositórios de dados cujas políticas correspondentes aceitam submissões de coleções de dados de várias áreas do conhecimento e que sejam provenientes de diferentes instituições de pesquisa.

Dois exemplos são essenciais para caracterizar os repositórios de dados multidisciplinares: o Dryad[7] e o Figshare[8].

O Dryad tem como objetivo arquivar dados de pesquisa que estão subjacentes a pesquisas descritas em publicações acadêmicas e vinculá-los a essas publicações. Por sua importância no cenário mundial de repositórios, ele é analisado mais detalhadamente na seção 6.

O FigsShare informa na sua página que ele “é um repositório onde os usuários podem tornar todos os seus produtos de pesquisa – incluindo coleções de dados, figuras e vídeos - disponíveis de maneira que possam ser citados, pesquisados e recuperados”. O Figshare permite imediata pré-publicação por meio de um portal web, e considera importante resultados negativos ou resultados que não seriam publicados de outra forma (THE

[7] http://datadryad.org/
[8] https://figshare.com/

ROYAL SOCIETY, 2012). Os dados podem ser compartilhados de forma restrita entre colaboradores ou tornados públicos em nome da pesquisa aberta, ou para se alinhar com as políticas mandatórias das agências de fomento. Todos os produtos de pesquisa tornados publicamente disponíveis recebem um DOI e são licenciados como *Creative Commons*. Este repositório iniciou suas atividades em 2011, e é operado pela Digital Science[9], companhia vinculada a Macmillan Publisher[10]. Apesar de ser um empreendimento privado, o *upload* de conteúdos e o acesso são gratuitos e de acordo com os princípios do movimento de acesso aberto.

Repositórios de dados de pesquisa orientados por projetos

São repositórios cujas coleções de dados são resultados de projetos de pesquisa ou resolução de problemas específicos. Um exemplo apontado por Pampel e seus colaboradores (2013) caracteriza bem essa categoria de repositório: The Scientific Drilling Database (SDDB)[11] que oferece dados de perfuração, abertos e reusáveis, que são criados no âmbito do Scientific Continental Drilling Program.

Esta classificação poderia se estender para coleções de dados armazenados em outros sistemas de informação e banco de dados fora do escopo científico, gerenciados por órgãos governamentais, empresas privadas e ONGs. Dependendo do nível de tratamento, da padronização de procedimentos e, sobretudo, da proveniência essas coleções de dados, se abertas, podem ser reusadas no ambiente de pesquisa gerando novos dados e novos conhecimentos.

A classificação apresentada pode ser útil para distinguir os possíveis tipos de repositórios de dados, no entanto ela não consegue espelhar os níveis de complexidade das infraestruturas que estão permeando os dados para garantir a permanência e a escala de acesso e de alcance desses dados. A

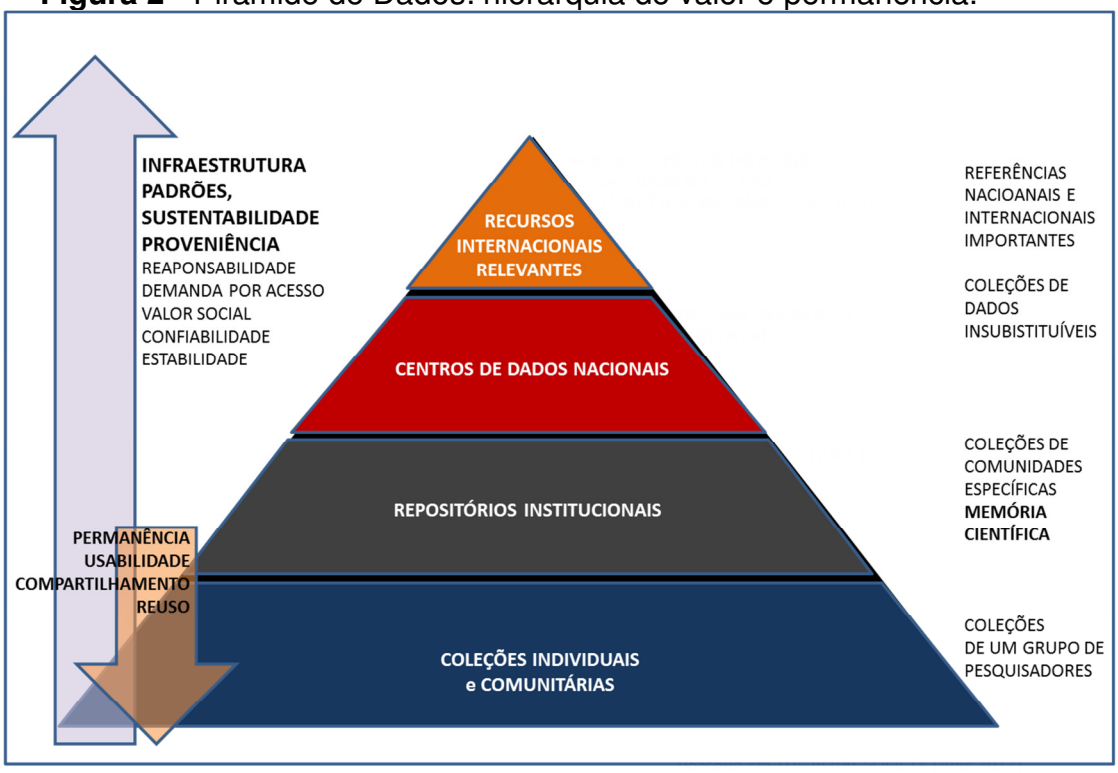
[9] https://www.digital-science.com/
[10] http://macmillan.com/
[11] http://gfzpublic.gfz-potsdam.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:236508

sistematização proposta pela The Royal Society (2012) deixa essa questão mais clara.

5 VALOR E PERMANÊNCIA DOS DADOS

O relatório da The Royal Society (2012) preconiza que há uma hierarquia de padrões de gestão de dados científicos que pode ser representado na forma de uma pirâmide de quatro camadas que, resumidamente, indicam o alcance dos sistemas: **internacional**, **nacional**, **institucional** e **comunitário** ou **pessoal**. As camadas refletem a disponibilidade para o acesso, os níveis de investimento, as infraestruturas necessárias, as reponsabilidades envolvidas, as ferramentas utilizadas e, sobretudo, a percepção da sua importância para ciência em escala planetária; a hierarquia reflete também os níveis de permanência dos dados e o grau de padronização de procedimentos que eles são submetidos. A figura 2 tenta sintetizar estas condições.

Figura 2 - Pirâmide de Dados: hierarquia de valor e permanência.



Fonte: Dos autores baseada em The Royal Society (2012)

- **CAMADA 1** – compreende os grandes **programas internacionais** que geram seus próprios dados, como acontece no Grande Colisor de Hádrons do CERN[12], que conta com uma rede complexa de computação em grade para distribuir seus dados mundialmente; inclui também os programas que fazem curadoria de dados provenientes de um grande número de fontes, como é, por exemplo, o Worldwilde Protein Data Bank[13] que se utiliza de ferramentas distribuídas para a submissão, curadoria e disseminação dos dados.
- **CAMADA 2** – inclui os centros de dados e demais infraestruturas informacionais voltados para dados de pesquisa gerenciados por organismos governamentais de âmbito **nacional** – como seria o caso do CNPq - ou por financiadores de projetos de pesquisa vinculados a instituições sem fins lucrativos. Muitos desses sistemas são gerenciados por instituições de pesquisa em nome de organismos governamentais nacionais.
- **CAMADA 3** – Processos de gestão e curadoria conduzidos por **instituições** individuais, tais como universidades e institutos de pesquisa para os dados gerados pelos seus próprios programas de pesquisa. Seus enfoques e metodologias variam muito. Embora muitos tenham políticas para gestão de dados de pesquisa, o documento tende a tomar forma de aconselhamento genérico para os pesquisadores, ao invés de ser uma base rigorosa e específica que esteja subjacente às atividades de curadoria e que considere todo o espectro de dados gerados pelas atividades de pesquisa da instituição.
- **CAMADA 4** – Conduzido por **pesquisadores individuais ou grupo de pesquisa** que estão, geralmente, fora de domínios disciplinares onde a norma é disponibilizar os dados em bases de dados internacionais, ou ande não existem bases de dados nacionais indicadas como repositórios de dados pelos financiadores de pesquisa. Neste caso os próprios pesquisadores precisam criar mecanismos para coletar e armazenar os seus dados, e o compartilhamento se restringe a colaboradores próximos e confiáveis; não obstante as limitações gerenciais e tecnológicas, as coleções de dados podem também estar acessíveis publicamente em *websites* institucionais ou do grupo ou projeto de pesquisa. O esforço de gestão e curadoria é

[12] http://home.cern/topics/large-hadron-collider
[13] http://www.wwpdb.org/

> geralmente financiado pelos agentes tradicionais, por meio de recursos financeiros aportados para projeto de pesquisa ou para o programa onde o projeto está vinculado.

As camadas não são estanques. Há uma dinâmica que faz com que as coleções de dados migrem para camadas superiores, principalmente quando dados de pequenas comunidades de pesquisadores se tornam de grande importância ou são absorvidas por outras bases de dados.

Essa circunstância impõe às bases de dados gerais uma condição de "incubadoras" de coleções de dados. Esse é caso da base de dados Dryad[14] – que analisaremos na próxima seção - que atua como um repositório para dados que necessitam estar disponíveis para fundamentar publicações e para cumprir as regras de disponibilização obrigatória de dados de um número crescente de periódicos. Essas coleções de dados não possuem repositórios específicos, mas à medida que se consolidam podem se estabelecerem como base de dados autônoma e especializada.

6 **DRYAD:** CONECTANDO DADOS ÀS PUBLICAÇÕES

A *homepage* do Repositório Digital Dryad informa que ele é um recurso informacional que opera sob um processo de curadoria que visa tornar os dados subjacentes às publicações científicas passiveis de serem descobertos, livremente reusados e citáveis. O Dryad se caracteriza como um repositório multipropósito para uma grande diversidade de tipos de dados de pesquisa, tendo como visão contribuir para um mundo onde os dados de pesquisa sejam disponíveis abertamente, integrados com a literatura acadêmica e rotineiramente reusados para criar conhecimento; e como missão oferecer uma infraestrutura para promover o reuso dos dados que fundamentam a literatura acadêmica e científica.

Conforme esclarece o portal, não obstante a sua característica multidisciplinar, o Dryad se originou a partir da iniciativa de um grupo vinculado

[14] http://datadryad.org/

a periódicos científicos relevantes e sociedades científicas que atuam nas áreas de Biologia Evolucionária e Ecologia. O primeiro passo foi na direção da adoção, para as suas publicações acadêmicas, de uma Política Conjunta de Arquivamento de Dados – mas conhecida pela sigla JDAP, Joint Data Archiving Policy[15]. A JDAP descreve os requisitos necessários para disponibilização pública de publicações que dão suporte a dados de pesquisa. O grupo pioneiro reconhece que a facilidade de uso, a sustentabilidade e uma infraestrutura governada pela comunidade são essenciais para apoiar essa política de disponibilidade.

Mesmo considerando que o JDAP e o Dryad são iniciativas distintas, via de regra, os periódicos que adotam JDAP recomendam Dryad como um repositório de dados apropriado para a curadoria dos dados associados aos itens de literatura científica que publicam.

Reconhecida a importância do Dryad, principalmente pela sua conexão próxima com os editores científicos e seu amplo alcance na qualidade de um repositório multidisciplinar, é necessário considerar, porém, que muitas vezes o pesquisador precisa depositar seus dados em repositórios mais específicos e disciplinares, que tenham políticas e procedimentos mais adequados para gestão de seus dados. Um problema que se coloca nesse momento é como encontrar, num universo de centenas de repositórios, um que atenda aos critérios exigidos. Essa é a questão que será examinada a seguir.

7 **RE3DATA.ORG:** TORNANDO OS REPOSITÓRIOS DE DADOS VISÍVEIS

O relatório da OECD (2007) preconiza como um ponto vital nas políticas nacionais de gestão de dados de pesquisa a necessidade de que os dados estejam visíveis e que possam ser facilmente descobertos. Significa dizer que as Informações sobre dados de pesquisa, sobre organizações que produzem dados, documentação sobre dados e especificações sobre as condições de reuso desses dados devem estar disponíveis internacionalmente de forma

[15] http://datadryad.org/pages/jdap

transparente, idealmente através da internet. “A falta de visibilidade dos recursos de dados de pesquisa existentes e de coleções futuras de dados coloca um série obstáculo ao acesso”, enfatiza o Relatório (OECD, 2007, p. 15). Porém, a crescente importância dos dados de pesquisa, somado às políticas mandatórias e de incentivos ao livre acesso e ao compartilhamento colocadas pelas agências de fomento e pelas instituições de pesquisa, somado às exigências de publicações dos dados por uma parcela considerável de editores de periódicos científicos, impulsiona o surgimento de serviços e sistemas em escala mundial voltados para hospedagem de dados científicos. Diante desse quadro se torna difícil para pesquisadores, agências financiadoras, editores e instituições acadêmicas selecionar repositórios apropriados para armazenar e descobrir dados de pesquisa.

Além do mais, a informação que é coletada e gerada nas atividades de pesquisa em vários domínios não é uniforme, pelo contrário, é heterogênea e de grande diversidade, isto significa que os dados têm sido publicados em uma infinidade de repositórios, e o número de novos repositórios continua crescendo (ULRICH et al., 2015).

Nesse panorama dominado pelo volume e heterogeneidade o Registry of Research Data Repositories Intiative[16] – mais conhecido pelo acrônimo “**re3data**.**org**” – desenvolveu e opera um diretório que reúne a descrição de repositórios digitais de dados de pesquisa em uma grande base de dados. “O alvo desse projeto é uma descrição indexada e estruturada dos RDR [repositórios de dados] de todos os domínios acumuladas num registro baseado em web” (PAMPEL et al., 2013, p. 7). Dessa forma, o diretório procura oferecer orientação aos pesquisadores no seu papel de produtores de dados bem como na qualidade de usuários de dados.

Uma particularidade importante das descrições dos repositórios apresentadas pelo **re3data**.**org** é um esquema de ícones que informam em um único olhar o conjunto de características mais importantes de cada repositório. Como assinala Pampel e seus colaboradores (2013), os ícones são

[16] http://www.re3data.org/

importantes não só para os pesquisadores, mas também para os operadores do sistema que podem comparar os repositórios e identificar fortalezas e fraqueza de suas próprias infraestruturas. O quadro 2 apresenta os ícones e seus significados, baseados em informações extraídas do *website* do diretório.

Quadro 2 – Ícones do re3data.org e seus significados

.

	ICONE	SIGNIFICADO
Informação		O repositório disponibiliza informações adicionais sobre os seus serviços
Acesso		O repositório oferece acesso aberto aos seus dados
		O repositório oferece acesso restrito aos seus dados
		O repositório oferece acesso fechado aos seus dados
Licenças		Os termos de uso e licenças dos dados são disponibilizados pelo repositório
Identificador Persistente		O repositório usa DOI para tornar seus dados persistentes, únicos e citáveis
		O repositório usa URN para tornar seus dados persistentes, únicos e citáveis
		O repositório usa ARK para tornar seus dados persistentes, únicos e citáveis
		O repositório usa HANDLES para tornar seus dados persistentes, únicos e citáveis
		O repositório usa PURL para tornar seus dados persistentes, únicos e citáveis
		O repositório usa outros esquemas de identificação para tornar seus dados persistentes, únicos e citáveis
Certificados e Padrões		O repositório é certificado ou segue os padrões para repositórios
Política		O repositório possui e disponibiliza um documento de política

Fonte: Autores, baseado em: Re3data (2016).

Na ausência de um esquema adequado de metadados, foi necessário desenvolver esquema próprio para descrever os repositórios. O vocabulário cobre os seguintes aspectos: informações gerais, responsabilidades, políticas, aspectos legais, padrões técnicos, padrões de qualidade e serviços (PAMPEL et al. 2013; VIERKANT et al., 2014).

O **re3data.org** pode ser considerado um sistema "baixa barreira de entrada". Entretanto, para um repositório ser indexado na sua base de dados alguns requisitos relacionados à disponibilidade de informações devem ser cumpridos. Detalhes sobre o modo de acesso e termos de uso dos dados são indispensáveis. Estas informações deviam estar naturalmente disponíveis, mas

apenas uma pequena parcela de repositórios tem políticas que explicitam as informações essenciais sobre seu serviço.

8 À GUISA DE CONCLUSÃO

A ciência digital terá como paradigma mais contundente o acesso aberto aos seus ativos informacionais que estão imbricados em todos os seus aparatos: instrumentos, modelos, códigos, metodologias, equipamentos e laboratórios. Um passo essencial nessa direção está expresso nas demandas da Ciência Aberta que incluem prioritariamente acesso livre aos produtos de pesquisas financiados com recursos públicos. Entretanto, os pressupostos desse movimento e a sua promoção vão exigir de várias instâncias – governamentais, financeiras, de pesquisa e ensino, éticas e legais – a instalação de uma infraestrutura permanente e sustentável de informações que estejam integradas às ciberinfraestruturas de pesquisa. Parte dessa infraestrutura esta resumida nas funções dos repositórios de dados que vão assistir os pesquisadores no compartilhamento de dados e na garantia dos princípios de reprodutibilidade e de autocorreção da ciência.

Nessa direção, repositórios de dados de pesquisa vão rapidamente se agregando, como um componente-chave, às redes mundiais de informação para a pesquisa. Nesse papel, tornam uma parte importante da atividade científica – antes oculta e sem lugar apropriado - visível e aberta para toda a sociedade; apoiam a validação e a revisão de publicações científicas; se tornam também parte da memória digital mais fidedigna da ciência, posto que podem registrar os percursos de erros e acertos que fazem parte do ciclo de geração de conhecimento científico; contribuem ainda para que os pesquisadores que coletam, geram e organizam dados possam ser identificados, reconhecidos e citados pelo trabalho que fica oculto no contexto de uma ciência voltada para o resultado final - a descoberta e a publicação. Entretanto é preciso assinalar que para os repositórios de dados cumprirem efetivamente seus objetivos é preciso que estejam orientados por uma política abrangente de gestão de dados de pesquisa que incluam vários outros elementos.

Há muitos desafios para a Ciência da Informação, Biblioteconomia e Ciência da Computação em torno dos repositórios digitais de dados de pesquisa, principalmente em termos de prática biblioteconômica e de aplicação de tecnologias. As bibliotecas de pesquisa – cumprindo sua missão ancestral - começam a expandir e renovar suas habilidades e conhecimentos, centradas em documentos, para estabelecer as bases de uma biblioteconomia de dados capaz de lidar com os estoques crescentes desses ativos informacionais (SALES; SAYÃO, 2015). Porém, os maiores desafios estão no âmbito da pesquisa interdisciplinar envolvendo especialmente as áreas acima citadas. Por exemplo, fazer com que essa diversidade de repositórios, com modelos de dados tão diferentes, consigam estabelecer níveis satisfatórios de interoperabilidade; desenvolver estratégias para a gestão por longo prazo de objetos digitais heterogêneos e complexos – que vão de simples séries numéricas a ambiente de realidade virtual –, o que é muito diferente do que gerenciar publicações digitais, já convencionais, como teses e artigos. Estes desafios tecnológicos e gerenciais vão redesenhando as bases de novas profissões como bibliotecários de dados, arquivistas de dados, cientistas de dados e especialistas em visualização de dados, mas, sobretudo, renovam as perspectivas de pesquisa para a Biblioteconomia e Ciência da Informação, integrando-as a patamares mais elevados, sem, entretanto, se afastarem dos seus mais preciosos fundamentos.

REFERÊNCIAS

RE3DATA. Registry of Research Data Repositories. **Re3data.org Reaches a Milestone & Begins Offering Badges**. Apr. 2016. Disponível em: <http://www.re3data.org/|>. Acesso em: 15 maio 2016.

BENSON, Dennis A. et al. GenBank. **Nucleic Acids Research**, Oxford, v. 41, Jan. 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531190/>. Acesso em: 15 maio 2016.

BORGMAN Christiane L. Research data: who will share what, with whom, when, and why? In: CHINA-NORTH AMERICAN LIBRARY CONFERENCE, 5., Beijing, 2010. **Proceedings...** Beijing: CALA, 2016. Disponível em: <http://works.bepress.com/borgman/238/>. Acesso em: 15 maio 2016.

CCSDS. Consultative Committee for Space Data System. **Reference model for an open archival information system (OAIS)**. Washington: Magenta book, CCSDS, 2002. Disponível em: <http://public.ccsds.org/publications/archive/650x0b1.pdf>. Acesso em: 15 maio 2016.

KINDLING, Maxi; SCHIRMBACHER, Peter. Die digitale forschungswelt als gegenstand der forschung. **Information**: Wissenschaft & Praxis, Berlin, v. 64, n. 2-3, p. 127–136, Apr. 2013.

KON, Fábio. Ciência aberta, dados abertos e código aberto. **Computação Brasil**, Porto Alegre, n. 22, jul. 2013. Disponível em: <http://www.ime.usp.br/~kon/papers/ComputacaoBrasilKon2013.pdf>. Acesso em: 15 maio 2016.

LEIDEN UNIVERSITY. **Publish and deposit your research**. 2015. Disponível em: <http://www.library.leiden.edu/teaching-researching-publishing/publish-deposit-research/data-management/>. Acesso em: 15 maio 2016.

NATIONAL SCIENCE BOARD. **Long-lived digital data collections**: enabling research and education in the 21st century. Arlington: National Science Foundation, 2005. Disponível em: <http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf>. Acesso em: 1 maio 2016.

OECD. Organisation for Economic Co-operation and Development. **OECD**: principles and guidelines for access to research data from public founding. 2007. Disponível em: <https://www.oecd.org/sti/sci-tech/38500813.pdf>. Acesso em: 15 maio 2016.

PANPEL, Heinz et al. Making research data repositories visible: the re3data.org Registry. **PLoS One**, San Francisco, v. 8, n.11, 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3817176/>. Acesso em: 15 maio 2016.

PÉREZ-GONZÁLEZ, Lourdes. Modelo/s de coste para la preservación de los datos científicos em la e-ciencia. In: JORNADAS DE GESTIÓN DE LA INFORMACIÓN, 12., 2010, Madrid. **Anales...** Madrid: SEDIC, 2010. Disponível em: <http://eprints.rclis.org/8555/1/Perez.pdf>. Acesso em: 1 maio 2016.

SALES, Luana Farias. **Integração semântica de publicações científicas e dados de pesquisa**: proposta de modelo de publicação ampliada para a área de Ciências Nucleares. 2014. Tese (Doutorado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

SALES, Luana Farias; SAYÃO, Luis Fernando. Há futuro para as bibliotecas de pesquisa no mundo da e-science. **Informação & Tecnologia**, João Pessoa, v. 2, n. 1, 2015. Disponível em: <http://periodicos.ufpb.br/ojs/index.php/itec/article/view/26029/14677>. Acesso em: 15 maio 2015.

SAYÃO, Luis Fernando; SALES, Luana Farias. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: CNEN, 2015. Disponível em: <http://carpedien.ien.gov.br:8080/handle/ien/1624>. Acesso em: 15 maio 2016.

THE ROYAL SOCIETY. **Science as an open enterprise**. London: The Royal Society Science Policy Centre, 2012. Disponível em: <https://royalsociety.org/~/media/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf>. Acesso em: 15 maio 2016.

ULRICH, Robert et al. **R3data.org**: making research data repositories visible and discoverable. 2015. Disponível em: <https://www.w3.org/2013/sharepsi/workshop/krems/papers/re3data>. Acesso em: 15 maio 2016.

UZWYSHYN, Ray. Research data repositories: the what, when, why, and how. **Computers in Libraries**, Westport, v. 36, n. 3, Apr. 2016. Disponível em: <http://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml>. Acesso em: 15 maio 2016.

VIERKANT, Paul et al. **Schema for the description of research data repositories**. 2014. Disponível em: <http://gfzpublic.gfzpotsdam.de/pubman/item/escidoc:758898:6/component/escidoc:775891/re3data_schema_v2-2_public_final-2014-12-03.pdf >. Acesso em: 15 maio 2016.

Title

Some considerations about digital research data repositories

Abstract
Introduction: Due to its commitment to seek new knowledge and new discoveries, contemporary scientific research produces and intensively uses digital research data. In this changing scenario, data are no longer simple byproducts of research activities, but have become first-rate information resources, characterizing a new scientific paradigm based on data sharing, access and reuse.
Objective: To identify the role of digital data repositories in the new scientific research scenarios, and to present an overview of their main characteristics, categories, benefits, functions and infrastructures.
Methodology: It analyzes the related literature and the main systems that give support to infrastructures for accessing and managing research data.

Results: In order for data collections to be able to transfer knowledge in a given time and space, as well as to be reused, it is quintessential to implement a permanent and sustainable technological and managerial infrastructure that allows these data collections to be curated throughout their life cycle. At the heart of this infrastructure are the digital repositories of research data, which are systems designed to support the selection, cataloging, archiving, access and sharing of research data.
Conclusion: Because of their importance as information resources, data repositories have quickly become an essential part of the research infrastructures on a global scale, by allowing an important part of the research activity visible and open to society as a whole. In this way, they pose some relevant challenges for Information and Library Science.

Keywords: Research data. Digital repository. Research data management. Digital curation. Open Science

Titulo

Algunas observaciones sobre los repositorios digitales de datos de investigación

Resumen
Introducción: La investigación científica contemporánea en su compromiso con la búsqueda de nuevos conocimientos y nuevos descubrimientos produce y hace uso intensivo de datos digitales de investigación. En este contexto de cambio, los datos ya no son simples subproductos de las actividades de investigación y se convierten en recursos de información de primera magnitud, que ofrece un nuevo paradigma científico guiado por el intercambio, amplio acceso y la reutilización de los dados de investigación.
Objetivo: Identificar el papel de los repositorios digitales de datos en el contexto de los nuevos escenarios de investigación científica y presentar una visión general de sus características principales, categorías, funciones, beneficios e infraestructuras.
Metodología: Se analiza la literatura del área y los principales sistemas que soportan la infraestructura de gestión de datos de acceso y la investigación.
Resultados: Este ensayo indica que para los datos y las colecciones de datos de investigación transmitan conocimiento en tiempo y espacio, y que pueden volver a utilizarse en otros contextos se requiere la implementación de una infraestructura tecnológica y de gestión, permanente y sostenible, que permite que los datos se mantiene durante todo su ciclo de vida. En el centro de esta infraestructura están repositorios digitales de datos de investigación, que son sistemas orientados a apoyar la selección, catalogación, archivamiento, acceso y el intercambio de los datos.
Conclusión: Debido a su importancia como una fuente de información, los repositorios de datos de investigación se convierten rápidamente en una parte esencial de las infraestructuras de investigación a escala mundial, haciendo visible y abierto a toda la sociedad una parte importante de la actividad científica. En este sentido, se convierten en un gran reto para la ciencia de la información y bibliotecología.

Palabras claves: Datos de investigación. Repositorio digital. Gestión de datos de investigación. Curaduría digital. Ciencia abierta.

Enviado em: 17.07.2016
Aceito em: 20.11.2016.