IX CONFERÊNCIA INTERNACIONAL SOBRE BIBLIOTECAS E REPOSITÓRIOS DIGITAIS DA AMÉRICA LATINA
BIREDIAL-ISTEC 2019
SÃO PAULO - BRASIL 30-31 DE JULHO / 1-2 DE AGOSTO 2019

A implantação da Base de Dados Científicos (BDC) da Universidade Federal do Paraná (UFPR)

Karolayne Costa Rodrigues de Lima[1], Paula Carina de Araújo[2], Janete Saldanha Bach Estevão[3]

[1] Bibliotecária na Universidade Federal do Paraná. Mestranda em Gestão da Informação pela Universidade Federal do Paraná (UFPR). Email: karolbraun@gmail.com
[2] Bibliotecária na Universidade Federal do Paraná. Doutora em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita (UNESP).
Email: paulacarina@ufpr.br
[3] Bibliotecária na Universidade Federal do Paraná. Doutoranda em Tecnologia pela Universidade Tecnológica Federal do Paraná (UTFPR). Email: janeteest@gmail.com

Resumo

A Base de Dados Científicos (BDC) da Universidade Federal do Paraná (UFPR) visa disponibilizar os dados científicos utilizados das pesquisas que são publicadas pela comunidade da UFPR em teses, dissertações, artigos de revistas e outros materiais bibliográficos. O objetivo desta comunicação é descrever o processo de implantação da BDC no âmbito da UFPR, desde as escolhas estratégicas até o seu efetivo uso. A implantação da BDC passou por algumas etapas como: a seleção do software, a seleção de requisitos, a seleção do padrão de metadados, o estabelecimento da metodologia para formular as diretrizes para o depósito dos dados, a especificação dos critérios para a criação do plano de gestão de dados, o registro no serviço indexador de repositórios de dados, o Re3data.org e a divulgação para a comunidade acadêmica. No decorrer do processo de planejamento e implantação da BDC, destacam-se alguns desafios: os custos com atribuição de DOI para cada *dataset*, o arranjo de características comuns para base de dados científicos que atendam aos contextos disciplinares distintos em uma única plataforma, a capacitação contínua dos usuários, a estrutura para oferta de serviços e a capacidade de suporte, além do entendimento do processo de produção e da evolução científica no contexto da ciência aberta.

Palavras-chave: Dados de pesquisa. Dados abertos. Repositório de dados. Base de Dados Científicos da Universidade Federal do Paraná (BDC/UFPR).

Abstract

The Scientific Database (CDB) of the Federal University of Paraná (UFPR) aims to make available the scientific data used from the researches that are published by the UFPR community in theses, dissertations, journal articles and other bibliographic materials. The thematic axis of this proposal is inserted in repositories of research data, specifically in the management and curation of data repositories. The purpose of this communication is to describe the process of implementation of the BDC within the scope of UFPR, from the strategic choices to its effective use. The implementation of BDC has undergone several steps: software selection, selection of requirements, selection of the metadata standard,

establishment of the methodology for formulating the guidelines for data storage, specification of the criteria for the creation of the plan data management, registration in the data repository index service, Re3data.org and dissemination to the academic community. During the process of planning and implementation of the BDC, some challenges are highlighted: the costs with DOI assignment for each dataset, the arrangement of common characteristics for scientific data bases that attend to the different disciplinary contexts in a single platform, the structure for offering services and support capacity, as well as the understanding of the production process and the scientific evolution in the context of open science.

Keywords: Scientific data. Open data. Data repository. Research data database of Paraná Federal University (BDC/UFPR).

Introdução

A partir dos anos 2000, as políticas que advogam pelo acesso aberto da comunicação da pesquisa científica, financiada com recursos públicos, têm impulsionado a difusão e o compartilhamento dos dados produzidos nos processos de investigação (CREASER, 2011, p. 56). Diversas instituições de fomento têm estabelecido políticas mais específicas de acesso aos dados de pesquisa daqueles que recebem recursos público (BERLIN, 2003; NATIONAL SCIENCE BOARD, 2005; NATIONAL SCIENCE FOUNDATION, 2007, 2010; OECD, 2007). Outra crescente exigência de depósito de dados tem sido protagonizada pelas revistas científicas, o que torna um pré-requisito aos pesquisadores estarem preparados para atuarem no cumprimento dessas regras. Um estudo sobre a atuação das revistas científicas no contexto de dados abertos mostrou que, dentre as 77 revistas brasileiras da área das ciências disponíveis no *Directory of Open Access Journals* (DOAJ), 15 delas já exigem o depósito de dados, sendo uma delas um periódico exclusivo de dados. Na área de Medicina, das 139 revistas indexadas, 71 fazem a mesma exigência (CARVALHO, 2016).

Buscando o contínuo alinhamento e a integração das melhores práticas em Ciência Aberta, a Universidade Federal do Paraná (UFPR), por meio de uma parceria entre o Centro de Computação Científica e Software Livre (C3SL) e o Sistema de Bibliotecas (SiBi) da UFPR, tendo a frente uma equipe multidisciplinar, planejou, desenhou e implementou, entre setembro de 2017 a janeiro de 2018, o primeiro repositório de uma universidade pública brasileira para dados científicos, a Base de Dados Científicos da Universidade Federal do Paraná (BDC/UFPR).

Portanto, o objetivo desta comunicação é descrever o processo de planejamento, decisões estratégicas e implantação da BDC/UFPR. Este projeto foi motivado pela contínua busca da excelência desta universidade em acompanhar a tendência mundial de planejamento, gestão, produção, organização, armazenamento, disseminação e reuso de dados de pesquisa. A disponibilização dos dados de pesquisa contribui para a transparência e otimização da produção científica por meio do reuso dos conjuntos de dados e a possibilidade de novas análises e abordagens.

Dados científicos de pesquisa

A mudança do paradigma científico provocada pela confluência das tecnologias de informação e comunicação no fazer científico tomou força a partir da metade para o final do

século XX e foi caracterizada por uma série de movimentos de abertura na prática e na comunicação da ciência. Como exemplo dessa abertura, aponta-se para o movimento do *software* livre de código aberto que permitiu o desenvolvimento de sistemas abertos (repositórios de dados); movimento de acesso aberto à informação científica na disponibilização por meio dos primeiros periódicos de acesso aberto; o fomento no desenvolvimento de recursos e práticas educacionais abertos, metodologia aberta, dados abertos, entre outros.

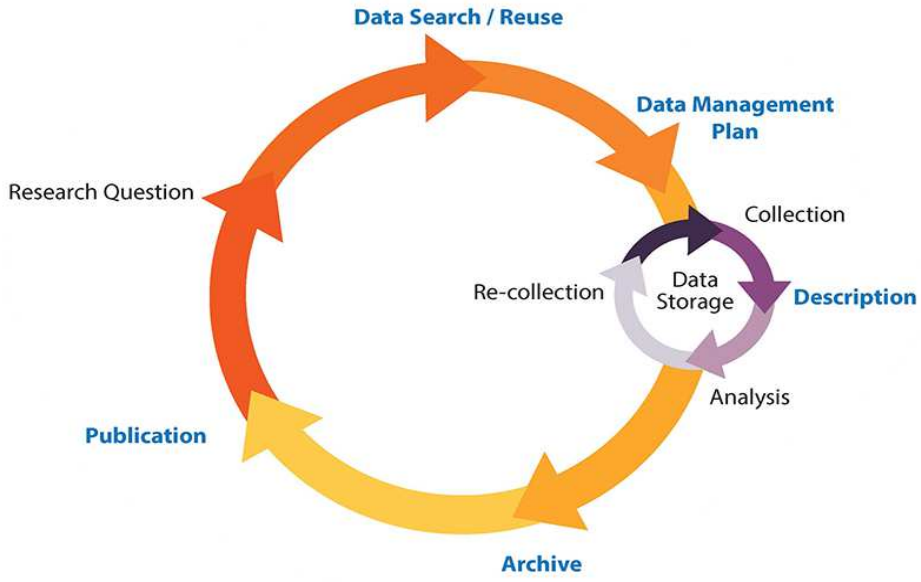
Nesse contexto de abertura, destacam-se ainda os dados científicos de pesquisa como uma unidade nuclear primordial do processo de investigação. Os dados científicos de pesquisa "referem-se ao material de fato registrado, comumente aceito na comunidade científica como necessário para validar os resultados da pesquisa" (THE ENGENEERING AND PHYSICAL SCIENCES RESEARCH COUNCIL, 2018, tradução nossa).

A considerar que o dado dentro de um projeto de pesquisa caracteriza-se como insumo ou produto, os conjuntos de dados (*datasets*) científicos têm características e particularidades dependendo de cada área do conhecimento e cobrem uma ampla gama de tipos de registros, podendo ser estruturados e armazenados em vários formatos de arquivos.

A pensar o contexto dos dados científicos de pesquisa dentro da perspectiva da comunicação científica, destacam-se certos princípios atrelados à abertura dos dados, os quais são: a **publicidade**, conferindo maior visibilidade aos pesquisadores; a indução da **colaboração** em rede e transparência dos dados utilizados para os resultados; a possibilidade de **reuso** de dados em novas conexões; a **aceleração** da produção de novas pesquisas; o atendimento às **regras de financiadoras** de pesquisa; a promoção da **reprodutividade**, a verificabilidade para garantir boa prática científica, evitando fraudes; e o **acesso** à pesquisa de importância social e maior consciência dos desafios da sociedade.

A pesquisa em torno dos dados científicos é ampla e não se restringe apenas ao seu armazenamento em repositório de dados que é o foco deste trabalho. Todavia, é relevante ilustrar o ciclo de gestão de um dado científico de pesquisa, conforme figura 1 abaixo:

Figura 1: Ciclo de gestão de dados científicos



Fonte: Digital Curation Center (2018).

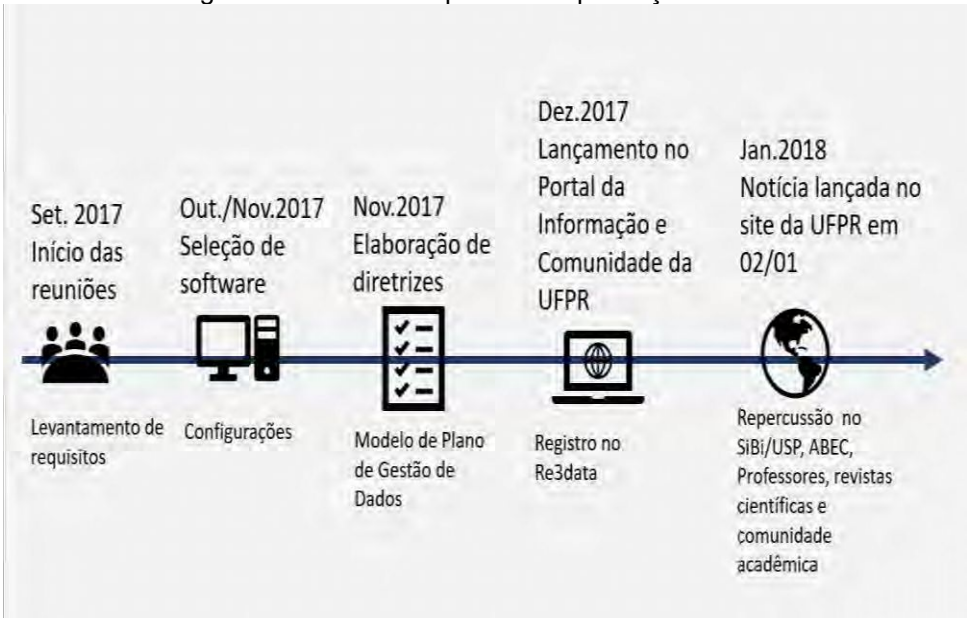
Base de Dados Científicos da Universidade Federal do Paraná (UFPR)

A Base de Dados Científicos (BDC) da Universidade Federal do Paraná (UFPR) foi lançada em dezembro de 2017 e é fruto do trabalho de uma equipe multidisciplinar do Centro de Computação Científica e Software Livre (C3SL) e o Sistema de Bibliotecas (SiBi) da UFPR. A BDC visa reunir os dados científicos utilizados nas pesquisas que foram publicadas pela comunidade da UFPR em teses, dissertações, artigos de revistas, e outros materiais bibliográficos. Trata-se de um repositório de caráter multidisciplinar para o depósito de dados científicos.

Além disso, a BDC é um serviço inovador que acompanha a tendência mundial de planejamento, gestão, produção, organização, armazenamento, disseminação e reuso de dados de pesquisa. A disponibilização dos dados de pesquisa contribui para a transparência e otimização da produção científica por meio do reuso dos conjuntos de dados e a possibilidade de novas análises e abordagens.

A implantação da BDC/UFPR ocorreu em diferentes etapas, tais como: a seleção do software; a seleção de requisitos; a seleção do padrão de metadados, o estabelecimento das metodologias para formular as diretrizes para o depósito dos dados, a especificação a seleção dos requisitos e dos critérios para a criação do plano de gestão de dados, o registro no serviço indexador de repositórios de dados, o Re3data.org e a divulgação para a comunidade acadêmica. Cada uma das etapas será descrita a seguir.

Figura 2: Linha do tempo até a implantação da BDC



Fonte: Setenareski; Estevão; Lima (2019)

- **Seleção do Software**: foram testados os serviços de hospedagem de dados e os *softwares* Zenodo, Fedora, Invenio, Open Science Framework e Ckan no estudo de viabilidade da ferramenta a ser utilizada. Porém, optou-se por manter o Dspace, *software* livre que com o qual a equipe está familiarizada, uma vez que já é usado para o Repositório Digital Institucional (RDI) da UFPR, e que atendeu aos pré-requisitos estabelecidos pela equipe, além das funcionalidades inerentes ao depósito de dados.

- **Seleção de Requisitos**: Foi utilizada a matriz de casos de uso e de requisitos funcionais, da Research Data Alliance (RDA)[1], selecionados para plataformas de repositório de dados de pesquisa (https://goo.gl/owqXHH).
- **Seleção de Padrão de Metadados**: o Diretório de Padrões de Metadados da RDA (https://goo.gl/DBjbvV) foi utilizado para um maior entendimento do que estava sendo adotado por outras instituições. Porém, dentre os repositórios registrados no Registry of Research Data Repositories (re3data.org[2]), o mais utilizado era o padrão Dublin Core, sendo este também adotado para a BDC/UFPR.
- **Metodologia para estabelecer as Diretrizes para depósito de dados**: a partir do método da análise de conteúdo, foi realizado um estudo das políticas para depósito de dados de instituições internacionais, que já estavam em estágio de maior maturidade em Repositórios de Dados Científicos: Revista Nature, Repositório Dryad, Repositório ICPSR, Repositório PLOS One, Repositório da Universidade de Edimburgo e Repositório Universidade do Texas. Também avaliou-se a Política do Inter-university Consortium for Political and Social Research (ICPSR[3]) e manteve-se o alinhamento com as orientações das recomendações do guia geral do Grupo de trabalho para políticas de repositório da RDA, o *Practical Policy Working Group*. As diretrizes apresentam os principais conceitos, orientações para elaboração do plano de gestão de dados; a estrutura e funcionamento da BDC/UFPR, incluindo orientações do processo de submissão e de direitos autorais.
- **Seleção de requisitos e formulação do Plano de Gestão de Dados**: o Plano de Gestão de Dados da BDC/UFPR foi resultado de uma avaliação de ferramentas automatizadas para geração de Plano de Gestão de Dados, tais como: - o DPMTool (https://dmptool.org); - o DCC (https://dmponline.dcc.ac.uk). Também foram avaliados modelos de Planos de Gestão de dados e outros padrões relevantes como: - o DMP Horizon 2020 (https://goo.gl/nfsEKf); - o DMP ICPSR (https://goo.gl/TyrNLZ). Essas ferramentas e modelos fundamentam as recomendações dos elementos principais que deveriam compor o Plano de Gestão de Dados da UFPR. O Plano de Gestão de Dados da BDC se propõe a descrever como os dados científicos serão tratados durante a pesquisa e, também, após a sua conclusão. É formado pelas seguintes seções: - identificação do pesquisador/grupo de pesquisa; - identificação da pesquisa; - descrição dos dados; - outras informações / metadados; - anuência das Diretrizes da BDC
- **Registro no Re3data.org**: prodeceu-se também o registro da BDC/UFPR no Registry of Research Data Repositories, principal indexador dos repositórios de dados em vários países.
- **Divulgação e Repercussão**: o lançamento da BDC/UFPR ocorreu no Portal da

[1] A RDA foi criada em 2013 com financiamento da Comissão Europeia, do NSF (Estados Unidos) e do departamento de inovação do governo australiano, com o objetivo de construir a infraestrutura social e técnica para permitir o compartilhamento aberto de dados científicos nas mais diversas áreas do conhecimento. Conta com mais de 5900 membros de 129 países (dados de agosto/17) (https://www.rd-alliance.org/node/51727). Apenas 2% dos membros da organização estão localizados na América Latina.

[2] A maior parte desses recursos indexados no re3data.org são repositórios para depósito de dados de instituições de pesquisas e de universidades. A pesquisa no re3data.org pode ser feita por um conjunto de critérios, como tipo de acesso, tipo de conteúdo, estrutura de dados, padrão de metadados utilizado, licença de uso de dados, país de origem, idioma, tipo de versionamento, área do conhecimento, *software* utilizados para o repositório, dentre outros.

[3] O ICPSR é consórcio internacional com mais de 750 instituições afiliadas no mundo que está na liderança em gestão de dados em Ciências Sociais, disponibilizando mais de 250 mil arquivos de dados científicos.

UFPR e houve a divulgação nas redes sociais. A Repercussão foi imediata, sendo compartilhado por Sistemas de Bibliotecas de outras universidades do país, despertando interesse também da Associação Brasileira de Editores Científicos (ABEC), de professores da universidade e externos a ela e de revistas científicas externas à UFPR. Houve grande interesse da comunidade acadêmica e, a equipe de suporte e orientação aos usuários interessados quando ao processo de depósito e documentação de *datasets* realizou 135 atendimentos apenas no primeiro semestre de 2018.

Após a implantação da BDC, há um grande esforço para a divulgação do serviço e esclarecimentos aos dos pesquisadores da UFPR quanto à importância e as vantagens de compartilharem seus dados de pesquisa desenvolvidas no âmbito da UFPR. A disponibilização dos dados se dá por meio da autossubmissão na comunidade específica dentro do repositório. A Figura 2 esclarece o processo de autossubmissão.

Figura 3: Autossubmissão de conjuntos de dados na BDC

Fonte: Setenareski; Estevão; Lima (2019)

A maioria dos atendimentos aos pesquisadores se deu por e-mail. Inicialmente, o pesquisador envia um e-mail à equipe da BDC solicitando a divulgação dos seus dados/arquivos na BDC. Em seguida, é enviado um e-mail ao pesquisar com um texto informativo sobre objetivo da BDC para que o mesmo informe se os dados/arquivos que tem interesse em publicizar na BDC são caracterizados como dados produzidos nos processos de investigação. Em caso positivo, o pesquisador precisará enviar um e-mail para a equipe da BDC com: nome completo, e-mail de contato, matrícula UFPR, vínculo com a UFPR (aluno, professor, outro), e informar a qual trabalho (artigo, tese, dissertação, monografia especialização/graduação) os seus dados/arquivos são referentes. Caso o pesquisador não tenha mais vínculo com a UFPR, informará à equipe da BDC: nome completo, e-mail de contato, matrícula UFPR, programa de (pós-) graduação da UFPR em que se formou, a qual trabalho (artigo, tese, dissertação, monografia especialização/graduação) os seus dados/arquivos são referentes, e a data de publicação/defesa deste trabalho.

Após a análise destas informações a equipe da BDC cadastra o pesquisador na BDC para que o mesmo proceda com a submissão dos seus dados/arquivos. O processo ocorre

por autosubmissão, sendo que o pesquisar recebe um tutorial que o auxilia no preenchimento do formulário de descrição dos dados, *upload* do *dataset*, documentação, licenciamento e disponibilização. Durante o processo de contato entre o pesquisar e a equipe da BDC é informado a possibilidade de criação de coleção específica para grupo de pesquisa da UFPR. Desta forma a BDC possibilita que mais de um pesquisador seja vinculado a esta coleção e possa submeter os dados/arquivos do grupo de pesquisa em um único lugar e em conjunto com os outros pesquisadores atuantes neste grupo de pesquisa.

Atualmente, há uma limitação de 2 gigabytes para o tamanho do *dataset* a ser depositado, dessa forma, *datasets* maiores são submetidos pela equipe da BDC. Também percebe-se que uma parte expressiva dos pesquisadores ainda não está habituada a elaborar um Plano de Gestão de Dados antes de iniciar a sua pesquisa, o que também representa um fator limitador. Compreende-se que é necessária uma mudança no processo de pesquisa no que diz respeito à disponibilidade dos dados, pois, ainda não há em todas as áreas do conhecimento, a ação do compartilhamento e do reuso dos dados de pesquisa. Em outros países, essa prática foi iniciada por força de leis e políticas institucionais que requerem o depósito dos dados ao final da pesquisa financiada com recursos públicos.

Durante os atendimentos, percebeu-se que muitos pesquisadores têm dificuldade em diferenciar a divulgação dos seus dados produzidos nos processos de investigação, da divulgação do trabalho ao qual estes dados estão vinculados. Este cenário, mostra a necessidade de capacitar os pesquisadores quanto ao entendimento da plataforma BDC. Para dirimir esta dificuldade, a equipe da BDC produziu uma Frequently Asked Questions (FAQ), em que detalha informações essenciais para o entendimento da BDC. Além disso, a equipe da BDC analisa os dados/arquivos submetidos na BDC e verificam se os mesmos estão adequados ao objetivo da plataforma. Em caso negativo entram em contato com o pesquisador para rever os dados submetidos.

Considerações Finais

Um projeto dessa natureza demanda o trabalho engajado em equipe, de pessoas de formações multidisciplinares. Bibliotecários, analistas e técnicos em Tecnologia da Informação e da Computação e professores contribuíram para a viabilidade deste serviço. Em termos de desafios, destacam-se os custos com atribuição de DOI para cada *dataset* e o arranjo de características comuns para base de dados científicos que atendam aos contextos disciplinares distintos em uma única plataforma. A capacitação contínua dos usuários, visto que é uma atividade recente no Brasil, principalmente em relação ao letramento no gerenciamento dos dados de pesquisa, também é um ponto de atenção. A estrutura para oferta de serviços e a capacidade de suporte, mantendo uma equipe multidisciplinar no atendimento constantemente atualizada, é um dos aspectos de maior impacto ao valor percebido dos serviços pela comunidade. Por fim, o entendimento do processo de produção científica e da evolução da Ciência no contexto do *e-Science*, são determinantes para o sucesso dos Repositórios de dados como um serviço aderente e relevante às necessidades dos pesquisadores do Século XXI.

Referências

AMORIM, Ricardo Carvalho; CASTRO, João Aguiar; SILVA, João Rocha da; RIBEIRO, Cristina. A Comparison of Research Data Management Platforms: Architecture, Flexible

Metadata and Interoperability. **Universal Access in the Information Society**, v. 16, n. 4, nov. 2017, p. 851–62. https://doi.org/10.1007/s10209-016- 0475-y. Acesso em: 19 mai. 2019.

ARAUJO, Alessandra Belezia; FERREIRA, Elisabete; FÜHR, Fabiane; MOREIRA, Fernando Cavalcanti; ESTEVÃO; Janete Saldanha Bach; LIMA, Karolayne Costa Rodrigues de; ARAÚJO, Paula Carina de; SETENARESKI, Ligia Eliana; GONÇALVES, Lucas Henrique; SCHMITZ, Rafaela Paula. **Diretrizes da Base de Dados Científicos da Universidade Federal do Paraná**. Curitiba, PR: SiBi/UFPR, 2018. Disponível em: https://www.portal.ufpr.br/documentos/BDC/diretrizes_BDC.pdf. Acesso em: 10 abr. 2019.

BERLIN Declaration on open access to knowledge in the sciences and humanities. Berlin, 2003. Disponível em: http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf. Acesso em: 08 mai. 2019.

CARVALHO, Teila Oliveira. A Influência das Revistas Científicas de Acesso Aberto para o Depósito e Publicação dos Dados de Pesquisa. In: 7. Conferência Luso-Brasileira sobre Acesso Aberto. **Anais**.... 2016. Disponível em: https://goo.gl/kjZhCM. Acesso em: 11 mai. 2019.

CREASER, Claire. Scholarly communication and access to research output. In: EVANS, WENDY, BAKER, DAVIS. **Libraries and Society**: role, responsibility and future in an age of change. EBSCO Publishing. 2011. p. 53-66.

DIGITAL CURATION CENTER. **Research data life cycle**. Disponível em: <http://www.dcc.ac.uk/>. Acesso em: 01 mai. 2019.

NATIONAL SCIENCE BOARD. **Long-lived digital data collections: enabling research and education in the 21st century**. National Science Foundation, Sept. 2005. Disponível em: http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf. Acesso em: 01 mai. 2019.

NATIONAL SCIENCE FOUNDATION. **Cyberinfrastructure Vision for 21st Century Discovery**. 2007. Disponível em: https://goo.gl/AEFEJ4. Acesso em: 10 abr. 2019.

NATIONAL SCIENCE FOUNDATION. **Dissemination and sharing of research results.** 2010. Disponível em: https://goo.gl/E6nsXV. Acesso em: 10 mai. 2019.

ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **OECD principles and guidelines for access to research data from public funding**. Paris: Organization for Economic Co-operation and Development, 2007. Disponível em: http://www.oecd.org/sti/sci-tech/38500813.pdf. Acesso em: 31 abr. 2019.

SETENARESKI, Ligia E.; Estevão, Janete Bach; Lima, Karolayne C. R. **Ciência aberta:** os movimentos que a impulsionam, como se relacionam e como se inserem no mercado das publicações científicas, 19 mai. 2019. 29 slides.