PLANO DE GERENCIAMENTO DE DADOS NO CONTEXTO DOS REPOSITÓRIOS DE DADOS DE UNIVERSIDADES

Data Management Plan in the context of university Data Repositories

Elizabete Cristina de Souza de Aguiar Monteiro[1], Ricardo Cesar Gonçalves Sant'ana [(2)]

(1) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, beteaguia@yahoo.com.br
(2) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, ricardosantana@marilia.unesp.br

Resumo:
Plano de Gerenciamento de Dados é o documento formal o qual descreve o conjunto de informações e instruções relacionado à gestão de dados científicos no seu ciclo de vida abordando critérios de coleta, organização, descrição, gerenciamento, disponibilização, acesso e curadoria dos dados tanto pelo pesquisador quanto pelos repositórios de dados. A elaboração do Plano de Gerenciamento de dados auxilia os pesquisadores e os profissionais atuantes nos repositórios. O objetivo deste estudo foi investigar quantos e quais repositórios de dados das 100 melhores universidades do mundo disponibilizam Planos de Gerenciamento de Dados e identificar aspectos relacionados a possíveis benefícios gerados pela adoção destes PGDs. A metodologia utilizada teve como base a pesquisa bibliográfica para a estruturação da fundamentação teórica concomitante à metodologia quantitativa e qualitativa. Foi utilizado o método exploratório para a coleta de dados para fazer o levantamento dos repositórios de dados das 100 melhores universidades do mundo através sítio *webometrics.* Os resultados mostram que, dentre as universidades alencadas, apenas 36 disponibilizam Planos de Gerenciamento de dados e que as instruções do PGD variam dependendo das características dos repositórios e dos conjuntos de dados neles depositados.
Palavras-chave: Plano de Gerenciamento de Dados; Gestão de dados; Dados científicos; Repositório de dados.

Abstract:
Data Management Plan is the formal document which describes the set of information and instructions related to the management of scientific data in its life cycle addressing criteria of collection, organization, description, management, availability, access and curation of the data both by the researcher and data repositories. The elaboration of the Data Management Plan helps the researchers and professionals working in the repositories. The purpose of this study was to investigate how many and which data repositories of the 100 best universities in the world provide Data Management Plans and identify aspects related to the possible benefits generated by the adoption of these PGDs. The methodology used was based on the bibliographical research for the structuring of the theoretical foundation concomitant to the quantitative and qualitative methodology. Was used the exploratory method to collect data to survey the data repositories of the 100 best universities in the world through website webometrics. The results show that, among the enlisted universities, only 36 provide Data Management Plans and that the instructions of the PGD vary depending on the characteristics of the repositories and the datasets deposited in them.
Keywords: Data Management Plan; Data management; Scientific data; Data repository.

1 Introdução

As instituições acadêmicas e científicas passam a ter cada vez mais a responsabilidade no gerenciamento de dados científicos coletados ou produzidos em grande quantidade, velocidade e variedade por pesquisadores nas diversas áreas do conhecimento. A gestão de dados requer, por parte de seus detentores, planejamento e ações concretas que tragam eficiência não só para coleta e armazenamento como também e, principalmente, para fase de

recuperação desses dados ampliando sua visibilidade e potencial uso (SANT'ANA, 2016).

Na ambiência da investigação científica, esse processo que agrega valor configura-se como importante fator para a ampliação do potencial de impacto dos resultados das pesquisas e da própria instituição (PRYOR, 2012)[1].

Desse contexto emerge a necessidade de políticas para o gerenciamento dos dados envolvidos nas pesquisas. Agências de fomento como a *National Science Foundation* (NSF), *National Institutes of Health* (NIH), *National Oceanographic Data Center* (NODC) e Nasa dos Estados Unidos, Horizon2020 da Europa, AHRC, BBSrc, *Cancer Research* UK, EPSRC, ESRC, MRC, NERC, STFC, *WELLCOMETrust* no Reino Unido estão incentivando, orientando ou mesmo tornando obrigatório a elaboração de Plano de Gerenciamento de Dados (PGD) para os projetos que terão o financiamento de suas pesquisas por essas agências (CORRÊA COUTO, 2016).

Um objetivo de destaque na gestão de dados científicos é assegurar que os mesmos possam ser compreendidos e interpretados por outros pesquisadores ao longo do tempo. Para isso é essencial uma descrição clara e detalhada dos dados, anotações adicionais e informações que contextualizam os dados e possibilitem que transmitam informação e conhecimento no tempo e no espaço (SAYÃO; SALES, 2015).

Os procedimentos descritos para a gestão de dados são documentados em um Plano de Gerenciamento de Dados (PGD), documento direcionado àqueles que estão envolvidos de alguma forma com a gestão desses dados (MONTEIRO, 2017). O PGD auxilia tanto os pesquisadores que coletam e manipulam conjuntos de dados quanto àqueles profissionais que atuam nos repositórios de dados científicos. Couto Corrêa (2016) complementa destacando que o PGD fornece diretrizes para todo o ciclo de vida dos dados.

O Plano de Gerenciamentos de Dados é um plano que descreve diferentes atividades e processos associados ao ciclo de vida de dados e envolve

> [...] a concepção e criação de dados, armazenamento, segurança, preservação, recuperação, partilha e reutilização, todos tendo em conta as capacidades técnicas, considerações éticas, questões legais e estruturas de governança. (COX; PINFIELD, 2014, tradução nossa).

Os procedimentos adotados na execução de um PGD definem e estabelecem métodos de execução das atividades e detalham os procedimentos que serão realizados. O planejamento é um processo cíclico, dinâmico e interativo, em que as fases não precisam ser lineares, pois há uma dinâmica no processo (ALMEIDA, 2005).

Conjuntos de dados de um determinado grupo de pesquisadores podem conter diferentes formatos, tipos e descrições, tornando-os altamente heterogêneos e, à medida que o tamanho dos conjuntos de dados aumenta, o seu gerenciamento tende a se tornar árduo (LEE et al., 2009).

Para contextualizar a coleta de dados, este trabalho utilizou o Ciclo de Vida dos Dados (CVD) (SANT'ANA, 2016), modelo composto por quatro fases: Coleta, Armazenamento, Recuperação e Descarte, sobre as

[1] conceito de Pryor retirado de informações referente ao livro PRYOR, G. (Ed.) **Managing research data**. United Kingdon: Facet Publishing, 2012. Disponível em: <http://www.dcc.ac.uk/news/book-managing-research-data>. Acesso em: 27 set. 2016.

quais perpassam por seis fatores como: Preservação, Disseminação, Direitos Autorais, Qualidade, Integração e Privacidade (Apêndice A).

A fase da coleta caracteriza o processo de obtenção dos dados. No contexto da coleta de dados científicos, participam diversos atores entre eles: pesquisador 1, detentor de dados (profissional responsável pelos dados no repositório), sociedade (pesquisadores 2, 3 e n que farão coleta nos repositórios).

A fase Coleta aparece tanto no momento do depósito dos dados no repositório pelo Pesquisador 1 (CVD - Repositório), quanto no momento da coleta dos pesquisadores 2, 3 e n (CVD - Pesquisador) quando coleta seus dados para sua pesquisa no repositório, conforme demonstrado no Apêndice B.

Desafios no âmbito da Ciência da Computação e da Ciência da Informação, tais como àqueles que ocorrem em todas as fases do CVD, Preservação, Disseminação, Direitos Autorais, Qualidade, Integração e Privacidade, permanecem em aberto o que torna difícil descobrir, compartilhar ou reutilizar dados pois:

1) dados valiosos podem ter sido descartados;
2) tecnologias da informação tendem a ter processo de obsolescência altamente acelerado;
3) formatos incompatíveis podem tornar os dados difíceis ou impossíveis de integrar;
4) o fluxo de dados entre domínios pode ser impedido por metadados incompletos, imprecisos e/ou mal descritos;
5) muitos cientistas relutam em compartilhar dados devido à falta de recompensa, às questões de propriedade intelectual e documentação apropriada (LEE et al., 2009).

Geralmente a instituição que implementa repositório de dados tem em seu sítio um documento ou informações relacionadas ao plano de gestão de dados para orientar os que depositarão conjuntos de dados na elaboração de PGDs. Cada fase e fator do Ciclo de Vida dos Dados devem ser considerados na elaboração do PGD.

Os Repositórios de dados científicos são ambientes digitais implementados nas universidades com infraestrutura para dar suporte aos pesquisadores no gerenciamento e na disponibilização de dados científicos facilitando a outros pesquisadores reutilizá-los (MONTEIRO, 2017).

Repositórios de dados científicos contribuem no gerenciamento de grandes quantidades de dados. Os repositórios de dados são mantidos por conjuntos de ações que viabilizam o armazenamento de dados visando à otimização da coleta pelos pesquisadores, o que amplia as potencialidades de reuso destes dados (MONTEIRO, 2017).

2 Objetivos

Os objetivos deste estudo foram investigar Repositórios de Dados das 100 melhores universidades do mundo para verificar quantos e quais disponibilizam em seus sítios Plano de Gerenciamento de Dados e identificar aspectos relacionados a possíveis benefícios gerados pela adoção destes PGDs.

3 Procedimentos Metodológicos

A metodologia utilizada teve como base a pesquisa bibliográfica concomitante à metodologia quantitativa e qualitativa. Foi utilizada a coleta de dados para o levantamento dos repositórios de dados das 100 melhores universidades do mundo. A coleta de dados se iniciou com a busca das melhores universidades do mundo por meio do ranking *webometrics.info.* definindo o escopo com as 100 melhores ranqueadas. A localização dos repositórios de dados nas universidades foi realizada nos meses de julho a setembro de 2016. Em seguida foi

realizada a pesquisa exploratória para o levantamento das páginas oficiais das universidades identificadas para localização dos repositórios de dados. Não foram analisados repositórios com acesso restrito ou com link quebrado. O processo de recuperação dos dados foi realizado por meio de coleta dos Planos de Gerenciamento de Dados dos repositórios de dados encontrados.

4 Resultados

A análise incluiu a identificação dos repositórios de dados das universidades e a identificação dos Planos de Gerenciamento de Dados.

O Apêndice C ilustra os caminhos que foram seguidos para a coleta nos repositórios de dados. Os resultados apontaram que 55 universidades dispõem de repositórios de dados. Dessas, 36 têm PGD, os quais forma analisados.

Os repositórios das universidades que têm PGD são: *Harvard University, Massachusetts Institute of Technology, Stanford University, University of California Berkeley*, *University of Michigan*, *University of Washington,* University of Wisconsin Madison*, University of Pennsylvania, University of Oxford, Yale University*, *University of Cambridge, Michigan State University*, *University of Texas Austin*, *University of California San Diego, Pennsylvania State University*, *University of Illinois Urbana Champaign*, *University of North Carolina Chapel Hill*, *Princeton University*, *University College London*, *University of British Columbia*, *University of Maryland Baltimore*, *Purdue University*, *California Institute of Technology Caltech, University of Virginia*, *University of California Irvine, University of Arizona*, *University of Edinburgh, Washington University Saint Louis*, *Simon Fraser University*, *Virginia Polytechnic Institute and State University*, *Tufts University, Ruprecht Karls Universität Heidelberg*, *University of Copenhagen*, *University of Amsterdam*, *Universiteit Utrecht*, *University of California Los Angeles UCLA.*

Os repositórios de dados documentam as instruções e normativas nos PGDs no qual mencionam vários aspectos do Ciclo de Vida dos Dados. As instruções inclusas no PGD variam dependendo das características dos repositórios e dos conjuntos de dados neles depositados.

Percebeu-se que cada repositório elaborou seu PGD de acordo com as necessidades e particularidades de sua comunidade e do tipo de conteúdo abordado nos conjuntos de dados.

Para auxiliar os pesquisadores na elaboração de Planos de Gerenciamento de Dados incluindo os requisitos necessários para tal, identificou-se o DMPtool[2] que é uma ferramenta da Universidade da Califórnia e que fornece orientações sobre instituições financiadoras específicas que exigem PGD e um guia para a elaboração do documento.

Os PGDs tem sua propriedade intelectual vinculada a quem os criou. O pesquisador que elabora o PGD no DMPtool pode optar em compartilhar seu PGD publicamente contribuindo com outros pesquisadores (DMPtoll, 2017).

Os usuários do DMPTool podem visualizar amostras de planos, requisitos das agências financiadoras e exibir as alterações mais recentes feitas em seus planos uma vez que permite ao usuário criar um documento editável para apresentar a uma agência de financiamento. Pode, ainda acomodar versões diferentes à medida que os requisitos de financiamento mudam (DMPtool, 2017).

O uso de PGD pelos pesquisadores e por repositórios de dados podem proporcionar benefícios a todos os envolvidos, pois orienta sobre vários aspectos conforme descrito nos vários repositórios pesquisados:

[2] https://dmptool.org/

- Fornece opções de acesso flexíveis: os dados ficam acessíveis a todos, ou com acesso restrito mediante solicitação, dependendo das opções do pesquisador;
- Os dados recebem um URL permanente sob a forma de um identificador de objeto digital (DOI) para que o pesquisador possa conectar seus dados para suas publicações;
- Acesso a longo prazo: identificadores persistentes e DOIs tornam mais fáceis aos pesquisadores localizarem e citarem os dados;
- Exemplos de como citar os conjuntos de dados fator que indica aspectos relacionados aos direitos autorais;
- Indicação de quais licenças estão atribuídas aos conjuntos de dados recomendando como os dados podem ser utilizados;
- Análise quantitativa: mostra com que frequência os dados são vistos e feito *download*;
- Maximiza a reutilização: pode ter consulta com o pesquisador para garantir que os dados estejam em um formato e estrutura que melhor facilite o acesso a longo prazo, descoberta e reutilização;
- Com o uso de padrão de metadados, os dados podem ser indexados pelo *Google*, o que aumenta a possibilidade de outros pesquisadores encontrarem os dados;
- O servidor possui serviço de *backups* e manutenção regular para evitar a perda de dados.
- Nos sítios contém indicações de ferramentas que auxiliam os autores a montarem seus PGDs;
- Indicações de quais licenças estão atribuídas aos conjuntos de dados, o que determina como os dados são licenciados e as formas de utilização.
- Pode auxiliar o pesquisador que vai depositar os dados atender aos requisitos das agências de fomento que se aplicam aos seus conjuntos de dados;
- Orienta sobre aspectos de privacidade dos dados, os quais devem ser anonimizados para que, quando compartilhados não ameaçarem a privacidade dos sujeitos referenciados.

Os procedimentos adotados na execução de um PGD são instrumentos que definem e estabelecem métodos de execução das atividades e detalham a forma exata pela qual os procedimentos serão realizados.

A preparação de um PGD envolve atividades em diferentes graus de formalidade, extensão, periodicidade, metas e objetivos. O desenvolvimento de seu conteúdo envolve as particularidades de cada área abrangida pelo repositório.

As instruções inclusas no PGD variam dependendo dos objetivos e das características dos repositórios e dos conjuntos de dados neles depositados. Essas instruções devem ser claras para não gerar dúvidas e inseguranças.

4 Considerações Finais

Plano de Gerenciamento de Dados é um documento que contribui para o desenvolvimento e gerenciamento dos dados nos repositórios com instruções aos pesquisadores no gerenciamento dos dados desde a coleta e aos profissionais que neles trabalham orientando no gerenciamento por meio de diretrizes para coleta, armazenamento, recuperação e descarte, contribuindo ainda, no atendimento de requisitos relacionados a privacidade, qualidade, integração, disseminação, direitos autorais e preservação dos conjuntos de dados.

Os resultados da análise demonstram que das 100 melhores universidades do mundo, apenas 36 delas disponibilizam PDGs. Quando

considerado que foram analisadas as melhores universidades do mundo, esse fator comprova que ainda se tem um longo caminho para a conscientização, por parte dos envolvidos na gestão dos dados, da importância do PGD para gerenciamento de dados científicos.

A gestão de dados é imprescindível para o bom andamento da pesquisa, porém, dentre as 55 universidades com Repositório de Dados, em 19 delas não foram localizados PGDs. As universidades implementaram o repositório de dados, no entanto, a gestão de dados ainda não está explicitamente evidenciada.

As vertentes apresentadas corroboram que a área da Ciência da Informação, por meio do seu arcabouço teórico e prático pode contribuir com a implementação de repositórios de dados ampliando a atenção dada à gestão de dados por meio do estudo e fomento da utilização de PGDs.

Referências

ALMEIDA, M. C. B. **Planejamento de bibliotecas e serviços de informação**. Brasília, DF: Brinquet de Lemos, 2005.

COX, A. M. PINFIELD, S. Research data management and libraries: current activities and future priorities. **Journal of Librarianship and Information Science**, London, v. 46, n. 4, p. 299-316, 2014. Disponível em: <http://lis.sagepub.com/content/46/4/299.full.pdf+html>. Acesso em: 27 set. 2016.

COUTO CORRÊA, F. Gestíon de datosde investigación. Barcelona: Editorial UOC, 2016. Disponível em: <http://bit.ly/2uwefAX>. Acesso em: 2 jul.2017.

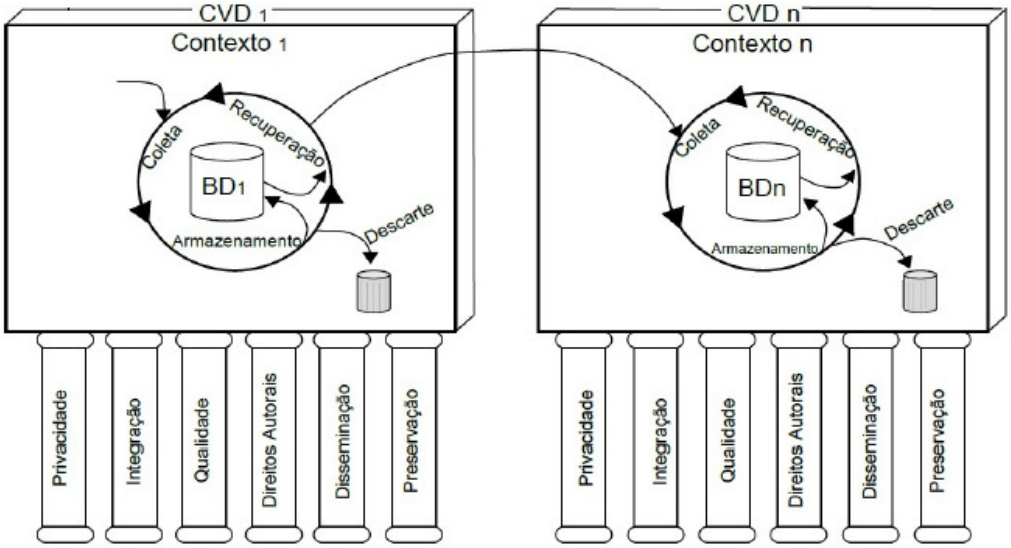
LEE, J. W. et al. DataNet: an emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning. **AIChe Journal,** New York, v. 55, n. 11, p. 2757-2764, Nov. 2009. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/aic.12085/epdf >. Acesso em: 05 jan. 2017.

MONTEIRO, E. C. S. A. **Direitos autorais nos repositórios de dados científicos: análise sobre os planos de gerenciamento dos dados.** 2017. 115 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de filosofia e Ciências, Universidade Estadual Paulista, Marília, 2017. Disponível em: <http://hdl.handle.net/11449/149748>. Acesso em: 30 abr. 2017.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação e informação**, Londrina, v. 21, n. 2, p. 116-142, maio/ago. 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940/20124 >. Acesso em: 20 out. 2016.

SAYÃO, L. F.; SALES. L. F. **Guia de gestão de dados de pesquisa para bibliotecários de pesquisadores**. Rio de Janeiro: CNEN, 2015. Disponível em: <http://carpedien.ien.gov.br:8080/bitstream/ien/1624/1/GUIA_DE_DADOS_DE_PESQUISA.pdf>. Acesso em: 5 out. 2016.

Apêndice A – Ciclo de Vida dos Dados para a Ciência da Informação (CVD-CI)

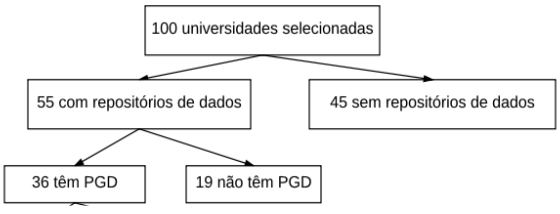


Fonte: SANT'ANA, 2016

Apêndice B – Ciclo de Vida dos Dados no Repositório

Fonte: MONTEIRO, 2017

Apêndice C – Direcionamento das análises



Fonte: Dados da pesquisa, 2017