Ricardo César Gonçalves Sant'Ana
Moisés Lima Dutra
Guilherme Ataíde Dias
Organizadores

ANAIS
WIDaT 2018

Organização do WIDaT 2018

- **Organização Geral:**

  Guilherme Ataíde Dias (PPGCI-UFPB) - Coordenador geral do evento

  Moisés Lima Dutra (PPGCIN-UFSC) - Vice-coordenador

- **Coordenador da Comissão Científica:**

  Ricardo César Gonçalves Sant'Ana (PPGCI-UNESP)

- **Comissão científica**

  Adilson Luiz Pinto (PPGCIN-UFSC)

  Ana Alice Baptista (Universidade do Minho, Portugal)

  Ana Carolina Simionato (PPGCI-UFSCar)

  Angela Maria Grossi de Carvalho (PPGCI-UNESP)

  Bernardina Maria Juvenal Freire de Oliveira (PPGCI-UFPB)

  Cristian Berrío-Zapata (PPGCI-UFPA)

  Dalton Lopes Martins (FCI-UnB)

  Denysson Axel Ribeiro Mota (PPGB-UFCA)

  Douglas Dyllon Jeronimo de Macedo (PPGCIN-UFSC)

  Ed Porto Bezerra (PPGI-UFPB)

  Edgar Bisset Alvarez (PPGCIN-UFSC)

  Edna Gusmão de Goés Brennand (MPGOA-UFPB)

  Edna Gomes Pinheiro (DCI-UFPB)

  Elaine Parra Affonso (FATEC-SP)

  Elvis Fusco (UNIVEM-Marília)

  Enrique Muriel Torrado (PPGCIN-UFSC)

  Evandro de Barros Costa (IC-UFAL)

  Fábio Paraguaçu (IC-UFAL)

  Fernando de Assis Rodrigues (PPGCI-UNESP)

  Gustavo Medeiros de Araújo (PPGCIN-UFSC)

  Henry Pôncio Cruz de Oliveira (PPGCI-UFPB)

  Joana Coeli Ribeiro Garcia (PPGCI-UFPB)

  José Eduardo Santarém Segundo (USP-FFCLRP)

  Leonardo Castro Botega (UNIVEM-Marília)

  Luana Farias Sales Marques (PPGCI-IBICT-UFRJ)

  Marckson Roberto Ferreira de Sousa (PPGCI-UFPB)

  Luís Fernando Sayão (CNEN)

  Marcelo Morandini (EACH-USP)

  Márcio Matias (PPGCIN-UFSC)

  Marcos Mucheroni (CBD-USP)

  Marynice de Medeiros Matos Autran (PPGCI-UFPB)

Maurício Barcellos Almeida (PPGGOC-UFMG)
Moisés Lima Dutra (PPGCIN-UFSC)
Plácida Leopoldina V. da Costa Santos (PPGCI-UNESP)
Pedro Luiz Pizzigatti Corrêa (POLI-USP)
Renata Baracho (PPGGOC-UFMG)
Ricardo César Gonçalves Sant'Ana (PPGCI-UNESP)
Robson Rodrigues Lemos (UFSC-Araranguá)
Rogério Ramalho (PPGCI-UFSCar)
Ryan Ribeiro de Azevedo (UFRPE-UAG)
Sandra de Albuquerque Siebra (PPGCI-UFPE)
Sandro Rautenberg (DECOMP-UNICENTRO)
Silvana Aparecida Borsetti G. Vidotti (PPGCI-UNESP)
Virginia Bentes Pinto (PPGCI-UFC)
Wagner Junqueira de Araújo (PPGCI-UFPB)
Zaira Regina Zafalon (PPGCI-UFSCar)

- **Coordenador do Cerimonial:**
  André Luiz Dias de França (PPGCI-UFPB)

- **Coordenador da Equipe Técnica Local:**
  Laerte Pereira da Silva Júnior (CCHLA-UFPB)

- **Equipe Técnica Local:**
  Adriana Alves Rodrigues (PPGCI-UFPB)
  Antonio Felipe dos Santos (MPGOA-UFPB)
  Débora Gomes de Araújo (PPGCI-UFPB)
  Pedro Augusto de Lima Barroso (PPGCI-UFPB)
  Pollianna Marys de Souza e Silva (PPGCI-UFPB)
  Renata Lemos dos Anjos (PPGCI-UFPB)

PRÁTICAS DE GESTÃO DE DADOS:
uma Revisão da Literatura Sobre o Termo *Data Life Cycle*

DATA MANAGEMENT PRACTICES:
a Review of the Term Data Life Cycle

Débora Gomes de Araújo[1]
(1) Universidade Federal da Paraíba (UFPB), Cidade Universitária, s/n - Castelo Branco III, João Pessoa - PB, 58051-085, debora.g.de.araujo@gmail.com.

Resumo:
Este trabalho realiza uma revisão bibliográfica sobre o termo "*data life cycle*", com enfoque na área da Ciência da Informação, tendo as bases de dados Emerald, LISA e LISTA como fontes de pesquisa. Nesse estudo foi utilizada uma abordagem quantitativa, de cunho exploratório e bibliográfico. Foram utilizadas as ferramentas Zotero e QDA Miner na condução da investigação. Quanto aos resultados, ficou evidenciado que a base de dados que mais indexou sobre a temática em questão foi a LISTA no período de 2013 a 2018, porém ainda apresentou uma quantidade pouco expressiva de indexações. O autor que se destacou na quantidade das produções científicas sobre o tema foi Vardigan, enquanto que o periódico que teve o maior número de publicação foi o americano *International Association for Social Science Information Service and Technology Quarterly* e o Reino Unido foi o local que mais publicou. Por meio dos artigos recuperados foi possível identificar a importância da gestão de dados em cada etapa do ciclo de vida de dados. Desta forma, este trabalho contribui para facilitar o entendimento da área pelos pesquisadores que investigam o tema.
Palavras-chaves: *Big data*. Ciência da Informação. Ciclo de vida dos dados. Gestão de dados científicos.

Abstract:
This work review the term "data life cycle" in the literature focused in the field of Information Science. Emerald, LISA and LISTA databases were used as sources of research. The study applied a quantitative approach with exploratory and bibliographic efforts. Zotero and QDA Miner tools were used to support the research. Although a little expressive amount of indexations have been found from 2013 to 2018, the results evidenced LISTA as the most indexed database. In addition, the recovered papers highlighted Vardigan as the most productive author in the field. The results also evidenced the American International Association for Social Science Information Service and Technology Quarterly as the most published periodical and the United Kingdom as the most published place. Through the retrieved papers it was possible to identify the importance of data management in each step of the data life cycle. This work will contribute to understanding the subject by the researchers in the field.
Keywords: Big data. Information Science. Data life cycle. Scientific data management.

1 INTRODUÇÃO

Nos dias atuais, cada vez mais a gestão de dados gerados a partir das pesquisas científicas vem ganhando espaço, devido as necessidades de uso e reuso dos mesmos. O que é facilitado pelas tecnologias contemporâneas. Neste sentido, com a disseminação da tecnologia da informação e comunicação (TIC) em nosso cotidiano, a pesquisa científica também evoluiu, em um contexto relacionado com o uso intenso dos dados, através da obtenção de informações de grandes volumes de dados digitalizados (AYDINOGLU; DOGAN; TASKIN, 2017).

O estudo em questão tem o seu foco nos dados de pesquisa, que de acordo com Patel (2016) estes constituem o centro de qualquer investigação científica, pois as descobertas e conclusões dos estudos são totalmente dependentes deles.

Dados de pesquisa incluem-se no processo dinâmico de investigação científica. Neste cenário, eles podem ser distinguidos por meio de duas funções: dados como materiais (*input*), está relacionada com a fase inicial, momento em que os dados são coletados e analisados e os dados como resultado (*output*), refere-se aos dados que são produzidos no decorrer dos processos, levando a publicações, como finalidade da pesquisa (SCHÖPFEL et al., 2016).

As práticas de pesquisas atuais exigem novas formas de tratamento dos dados por parte dos pesquisadores, de forma a acompanhar as mudanças constantes. Diversas áreas do conhecimento desenvolvem planos de gestão de dados por exigências de agências fomentadoras de pesquisa. O ciclo de vida de dados (CVD) se apresenta nesse cenário como uma ferramenta que pode promover habilidades e conhecimento para o pesquisador conduzir de forma apropriada a sua pesquisa, oferecendo etapas que contemple todo o percurso dos dados, de forma a serem detectáveis e utilizáveis em outros estudos.

Desta forma, é fundamental compreender o CVD, pois além de ser essencial em seu próprio trabalho, contribui com os pesquisadores, ao possibilitar soluções para as barreiras que podem ser encontradas na coleta e análise de dados, assim como na organização de conjuntos de dados e na descoberta de conjuntos externos relevantes (GOBEN; RASZEWSKI, 2015).

Segundo Rice e Southall (2016), um ciclo de vida de pesquisa mapeia a atividade de um pesquisador durante um projeto de pesquisa. Analogamente, um CVD representa o percurso dos dados ou as ações necessárias para que a pesquisa evolua para o estágio seguinte.

Na literatura existem algumas iniciativas que representam os ciclos de vida dos dados. De acordo com Goben e Raszewski (2015), existem vários modelos capazes de descrevê-los, porém eles citam como exemplos mais populares o *Digital Curation Center* (DCC) e a iniciativa do *Data Observation Network for Earth* (DataONE). Outras iniciativas podem ser evidenciadas como o *Data Documentation Initiative* (DDI), apresentada por Vardigan (2013), assim como os modelos destacados por Rice e Southall (2016), da *Joint Information Systems Committee* (Jisc) (2016) e do *Imperial College London Library RDM Workflow*, cujo autor é Barnes (2016), existem ainda diversos modelos de ciclo de vida de dados.

Da realidade da coleta, armazenamento, recuperação e descarte de dados, Santa'Ana (2016) desenvolveu um modelo voltado para a ciência da informação (CI), uma vez que, enfatiza o papel desta área no estudo e na proposta de caminhos para lidar adequadamente com os dados, ao revelar que a área pode trabalhar com um novo enfoque, sendo uma aliada nesse processo de otimização de dados.

2 OBJETIVOS

Diante destas considerações iniciais, o objetivo da presente pesquisa foi realizar uma revisão bibliográfica sobre o termo *data life cycle*, na literatura concentrada na área da Ciência da Informação (CI), nas bases de dados *Emerald eJournals Premier* (Emerald), *Library and Information Science Abstracts* (LISA) e *Library e Information Science & Technology Abstracts* (LISTA). Uma vez que, estas bases registram de formas variadas e focam de forma específica na CI.

3 PROCEDIMENTOS METODOLÓGICOS

O estudo em questão é de cunho quantitativo, exploratório e bibliográfico. Teve como fonte de dados os metadados dos artigos de arquivos abertos encontrados nas bases de dados supracitadas. A fase de coleta compreendeu o período de 01/07/2018 a 23/07/2018. Foram adotados os seguintes critérios de busca: utilizou-se apenas o termo "*data life cycle*", o qual foi coletado entre aspas, o espaço temporal de levantamento foi entre 2013-2018, com a finalidade de identificar o que vem sendo trabalhado sobre o tema nos últimos cinco anos.

Os dados revelaram que a base que mais indexou artigos de arquivos abertos sobre o *data life cycle* no período em questão foi a LISTA, pois através dela foi possível recuperar 35 artigos. Sendo 22 provenientes da Emerald e 14 da LISA. Em 2015 houve um avanço nas produções, o que pôde ser verificado nas três bases. Em 2016 o volume foi significativo. Apresentou uma queda em 2017 na base LISTA, pois não houve indexações. Em 2018 as referidas bases continuam disponibilizando artigos sobre a temática. Desta forma, o tema é alvo de discussão no cenário atual.

Os artigos oriundos da busca bibliográfica foram exportados para a ferramenta ZOTERO, através desta foi possível constatar na opção itens duplicados, os artigos recuperados que se repetiam entre as bases de dados. Através dessa análise, verificou-se que dos 14 artigos encontrados na base de dados LISA, 12 deles estão presentes nas bases de dados Emerald e LISTA, sendo 10 na Emerald e 2 na LISTA. Em seguida, os registros duplicados foram eliminados, passando do total de 71 para 59 artigos.

4 RESULTADOS

A partir do levantamento dos autores que produziram sobre o tema, constatou-se que apenas 13% deles publicaram mais de uma vez sobre o termo *data life cycle*. Enquanto uma quantidade significativa de autores, representada por 87%, discorreu pouco sobre o tema, pois apresentaram apenas um artigo. Infere-se que uma quantidade reduzida de estudiosos produziu um pouco mais sobre a temática, ao passo que um número considerável de pesquisadores está publicando pouco.

A autora que se destacou na quantidade de artigos sobre o tema foi Mary Vardigan, produziu 4 artigos nos períodos de 2013 a 2016, no periódico americano *IASSIST Quarterly*, sendo que a

produção de 2013 foi apenas de sua autoria e as demais envolveram outros pesquisadores, sendo duas delas com Ionescu.

Constatou-se que os autores Borgman, Darch, Sands, Wallis e Traweek produziram dois artigos juntos. Os pesquisadores Schöpfel, Južnič, Prost, Malleret, Češarek e Koler-Povh em conjunto produziram dois artigos e os autores Si, Xing, Zhuang e Hua também tiveram dois trabalhos juntos. Desta forma, a maioria dos estudiosos que apresentaram mais de um artigo tiveram seus trabalhos em parceria com outros que se enquadraram nas mesmas condições, com exceção de Harris que publicou sozinho dois artigos e Li que desenvolveu suas pesquisas com outros pesquisadores que tiveram apenas 1 publicação.

A análise possibilitou levantar os periódicos que apresentaram mais produções no período sobre a temática. Destacamos os que têm acima de uma publicação, conforme o Quadro 1.

Quadro 1: Distribuição dos periódicos por países no período de 2013-2018.

TÍTULO DE PUBLICAÇÃO	NÚMERO	PAÍSES
IASSIST Quarterly	8	Estados Unidos
International Journal on Digital Libraries	3	Alemanha
Information Services & Use	3	Holanda
Journal of Map & Geography Libraries	3	Estados Unidos
Information & Computer Security	2	Inglaterra
Journal of The Association for Information Science and Technology	2	Estados Unidos
Library Hi Tech	2	Inglaterra
New Review of Information Networking	2	Inglaterra
Program: electronic library and information systems	2	Inglaterra
The Electronic Library	2	Inglaterra
The Canadian Journal of Library and Information Practice and Research	2	Canadá
The Grey Journal	2	Holanda

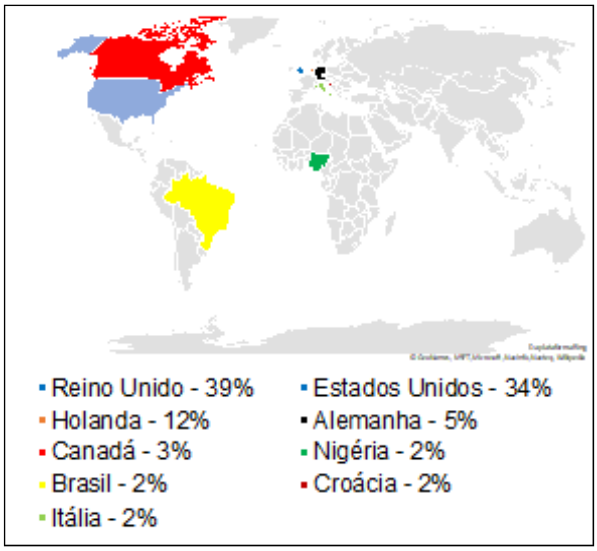
Fonte: Dados da pesquisa, 2018.

Através do levantamento dos Títulos de publicações, foi possível constatar que o maior número de publicações ocorreu no periódico americano *IASSIST*[1] *Quarterly*, ao apresentar 8 produções.

Ao incluir os artigos de outros periódicos que tiveram apenas uma produção, foi possível verificar os países que trataram da temática. O que está expresso na Figura 1.

1 International Association for Social Science Information Service and Technology

Figura 1: Países que publicaram sobre *data life cycle* no período de 2013-2018.



Dados da pesquisa, 2018.

Diante dos dados apresentados, é possível identificar que o *data life cycle* é um tema discutido no cenário nacional e internacional, com uma concentração de publicações em periódicos do continente europeu, representando 60%, o que pode ser resultado de iniciativas como o Horizon 2020 (H2020), que consiste em um programa de pesquisa e inovação que tem por finalidade aperfeiçoar o acesso à informação científica, no que tange aos artigos de pesquisa científica e aos dados de pesquisa. Em um contexto digital de Open Acess (HORIZON, 2018).

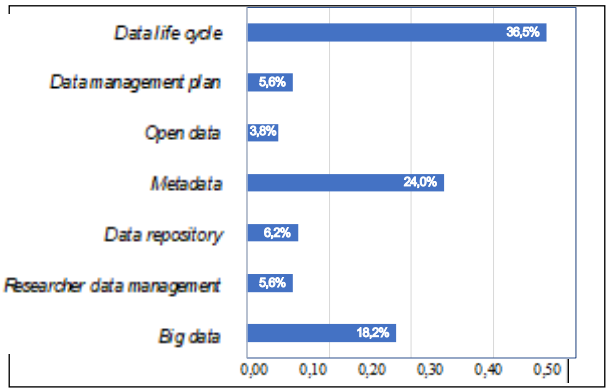
Nesse cenário, observa-se que o Reino Unido se destacou com um percentual de 39%, o que pode remeter a ser um local de referência na criação de ciclos de vida de dados, como por exemplo, o DCC, a iniciativa da Jisc e a do *Imperial College London Library RDM Workflow.* Além dessas estratégias voltadas para lidar corretamente com os dados de pesquisas, Hua et al. (2015) destacam que a biblioteca da Universidade de Cambridge oferece treinamentos para os pesquisadores gerenciarem seus dados.

Posteriormente, aparece o continente americano com 39% das produções, sendo 34% dos Estados Unidos. Diante desta realidade, constata-se a presença de iniciativas que visam a gestão de dados no cenário americano, como a criação dos ciclos de vida de dados DataOne e o DDI. Com uma participação reduzida de 2%, encontramos também uma produção brasileira, que também apresenta um modelo de ciclo de vida de dados com o foco na CI, do autor Sant'Ana (2016).

O estudo revela que apesar do autor com maior número de publicações e o periódico que mais se destacou serem americanos, o local que mais produziu atualmente sobre o data life cycle é o Reino Unido, seguido dos Estados Unidos.

Tomando como base as palavras-chave dos artigos citados pelos autores que publicaram em mais de um artigo, por meio da ferramenta QDA MINER, foi realizada uma pesquisa, a partir da criação de códigos nomeados pelas palavras *data life cycle* (ciclo de vida de dados), big data, cujo termo se refere a volume, variedade, velocidade e veracidade dos dados, segundo Dale (2015), *research data management* (gestão de dados de pesquisa), *data repositor* (repositórios de dados), *metadata* (metadados), open data (acesso aberto) e *data management plan* (plano de gestão de dados), para verificar a frequência desses termos em todos os arquivos recuperados, exceto os que estavam presentes nas referências dos artigos analisados (Gráfico 1).

Gráfico 1: Frequência dos termos encontrados nos artigos sobre *data life cycle* utilizando a ferramenta QDA Miner.

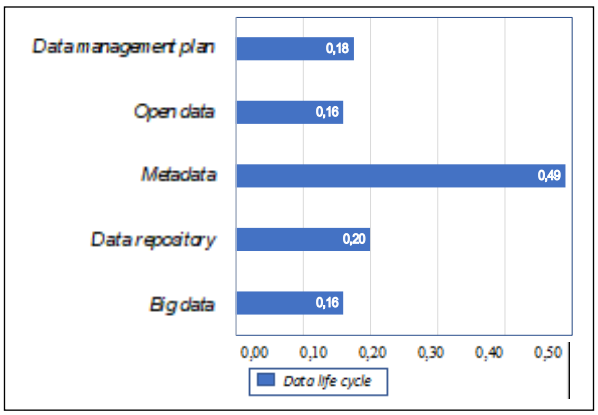


Fonte: Dados da pesquisa, 2018.

O Gráfico 1 confirma que o termo mais frequente nos textos avaliados foi o *data life cycle*, haja vista que os arquivos recuperados nas bases de dados se basearam no referido termo. Os demais termos estão envolvidos com a temática com destaque para *metadata* com 24% e *big data* com 18,2%. Por meio desse levantamento foi possível codificar os textos com as palavras supramencionadas, possibilitando a extração de segmentos relevantes contidos nos textos para o apoio teórico da pesquisa.

O Gráfico 2 a seguir, apresenta a coocorrência dos termos com relação ao *data life cycle*, ou seja, as vezes que as citações dos termos ocorreram simultaneamente, o que revela a proximidade dos mesmos.

Gráfico 2: Proximidade dos termos encontrados nos artigos sobre *data life cycle* utilizando a ferramenta QDA Miner.



Fonte: Dados da pesquisa, 2018

Apesar do *big data* ter sido o termo mais citado nos textos recuperados pelo Gráfico 1, ele juntamente com o *open data* são os mais distantes do termo *data life cycle* na análise de proximidade (Gráfico 2) e o *Researcher data management* não apresentou coocorrência. O termo *metadata* foi recuperado como o mais próximo, seguido do *data repository*. Essa análise se mostra importante quando na recuperação dos termos coocorrentes, o que ajuda na identificação de termos simultâneos.

5 CONSIDERAÇÕES FINAIS

Através da pesquisa realizada foi possível fazer uma revisão da literatura sobre o termo *data life cycle*, nas bases de dados Emerald, LISA e LISTA. Evidenciou-se que a temática ainda é emergente na área da Ciência da Informação, uma vez que, a quantidade de produções não foi expressiva no período analisado. No entanto, a *IASSIST QUARTERLY* se destacou pela quantidade de publicações na área da CI, denotando alinhamento com os interesses dos profissionais da área sobre a temática, mesmo que ainda seja pouco explorada.

O termo é discutido no cenário internacional, com destaque para o Reino Unido, visto o maior de número de produções. Mary Vardigan, que mais produziu no período estudado, trabalha em projetos relacionados à administração de dados, tendo atuado como diretora do DDI, que como dito anteriormente, representa um ciclo de vida dos dados, podendo ser considerada uma referência no assunto.

As ferramentas Zotero e QDA Miner ofereceram um apoio na condução da pesquisa. Através da segunda foi possível extrair dos textos citações relevantes para o trabalho, facilitando o processo de pesquisa, em que a partir dos segmentos selecionados ficou claro que o processo de gerenciamento de dados científicos está diretamente ligado com o ciclo de vida dos dados, sendo tal conexão algo fundamental para possibilitar o compartilhamento dos dados, uma vez que, boas práticas de gerenciamento são necessárias em cada fase do ciclo de vida dos dados.

Recomendamos como sugestão de trabalhos futuros, o desenvolvimento de pesquisas que, além de fazer um levantamento dos autores que mais publicaram na área, mostrem a quantidade de citações que têm recebido. Sugerimos ainda que estudos sejam realizados para identificar os ciclos de vida de dados existentes no âmbito científico, empresarial e governamental.

REFERÊNCIAS

AYDINOGLU, A. U.; DOGAN, G.; TASKIN, Z. Research data management in Turkey: perceptions and practices. **Library Hi Tech**, v. 35, n. 2, p. 271–289, 27 abr. 2017. Disponível em< https://www.emeraldinsight.com/doi/abs/10.1108/LHT-11-2016-0134> Acesso em 17 de jul. 2018.

BARNES, A. Imperial Colegge London Library RDM Workflow. Zenodo. 2016. Disponível em:< https://zenodo.org/record/54000#.W6FzFmhKjIU>. Acesso em 18 set. 2018.

DALE, K. L. RIM's Role in Harnessing the Power of Big Data. **Information Management Journal**, v. 49, n. 4, p. 29–32, 7 ago. 2015.

DATAONE. **Primer on Data Management**: what you always wanted know. Disponível em: <https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf>. Acesso em: 20 jun.2018.

DCC Digital Curation Centre. **Curation Lifecycle Model**. Disponível em: <http://www.dcc.ac.uk/resources/curation-lifecycle-model> Acesso em: 20 jun. 2018.

GOBEN, A.; RASZEWSKI, R. The data life cycle applied to our own data. **Journal of the Medical Library Association**, v. 103, n. 1, p. 40–44, jan. 2015. DOI: http://dx.doi.org/10.3163/1536-5050.103.1.008. Disponível em< https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4279933/> Acesso em 17 de jul. 2018.

H2020. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Disponível em:<http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf>. Acesso em 8 ago. 2018.

HUA, X. et al. Investigation and analysis of research data services in university libraries. **The Electronic Library**, v. 33, n. 3, p. 417–449, 27 maio 2015.

Joint Information Systems Committee.2016. Disponível em< https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data> Acesso em: 28 set. 2018.

PATEL, D. Research data management: a conceptual framework. **Library Review**, v. 65, n. 4/5, p. 226–241, maio 2016.

QDA MINER. Disponível em< https://provalisresearch.com/products/qualitative-data-analysis-software/> Acesso em: 25 jul. 2018.

RICE, R.; SOUTHALL, J. **The Data Librarian´s handbook.** Publisher by Facet Publishing. London, 2016.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, v. 21, n. 2, p. 116, 20 dez. 2016.

SCHÖPFEL, J. et al. Dissertations and Data. **Grey Journal (TGJ)**, v. 12, n. 3, p. 126–148, set. 2016.

VARDIGAN, M. The DDI Matures: 1997 to the Present. **IASSIST Quarterly**, v. 37, n. 1–4, p. 45–50, mar. 2013.

ZOTERO. Disponível em< https://www.zotero.org/> Acesso em 27 jul. 2018.