mepad

Métodos e Pesquisa em Administração
v. 4, n. 1, p. 35-47, 2019
Seção: Pesquisas científicas

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
Universidade Federal da Paraíba
debora.g.de.araujo@gmail.com

Guilherme Ataíde Dias
Universidade Federal da Paraíba

Wagner Junqueira de Araújo
Universidade Federal da Paraíba

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, o gerenciamento de dados gerados das pesquisas científicas vem ganhando espaço devido as necessidades de socialização 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).

Neste cenário, temas como o *Big data* foram surgindo, criando novas oportunidades e novos problemas. O termo *Big data* refere-se à heterogeneidade dos dados, a sua rápida geração e a grande quantidade que é disponibilizada digitalmente (MAYER-SCHÖNBERGER; CUKIER, 2013). Os autores Coyne, Coyne e Walker (2018) mostram que as organizações enfrentam o desafio de lidar com um volume cada vez maior de dados e para enfrentar essa realidade estão buscando o armazenamento digital. Assim sendo, os dados constituem um assunto de grande relevância no âmbito governamental, empresarial e científico, não devendo ter o seu gerenciamento negligenciado.

Dale (2015) apresenta 4 Vs que estão relacionados com o *Big data*, os quais são volume, que se refere a quantidade de dados existentes, variedade, representa os locais e tipos de dados, velocidade, referente a agilidade que os dados são gerados e a veracidade que trata da qualidade dos mesmos. Segundo Federer (2016), o termo é comumente utilizado sobre os desafios do gerenciamento de dados relacionados à pesquisa, dados científicos.

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

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 já necessitam desenvolver planos de gestão de seus dados por exigências de agências fomentadoras de pesquisa. O ciclo de vida de dados se apresenta como uma ferramenta de gestão 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.

É fundamental compreender o ciclo de vida dos dados, pois além de ser essencial em seu próprio trabalho, contribui com os pesquisadores, possibilitando 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 de dados externos relevantes (GOBEN; RASZEWSKI, 2015). Este fato contribui com o compartilhamento dos dados, por isso suas etapas devem ser esclarecidas, para que os dados frutos de outras investigações possam contribuir com novas pesquisas.

Diante desta realidade de uso, acesso e manutenção dos dados de pesquisa, Santa'Ana (2016) enfatiza o papel da Ciência da Informação (CI) 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.

A partir 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, para isso foram consultadas as 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.

2 CICLO DE VIDA DOS DADOS

O papel dos dados é de suma importância para qualquer pesquisa ou projeto de pesquisa. Eles são cuidadosamente coletados, analisados, otimizados, organizados, de forma a serem úteis para a realização de estudos. As investigações seriam comprometidas sem dados autênticos e objetivos (PATEL, 2016).

De acordo com Schöpfel et al. (2016), os dados de pesquisa fazem parte do processo dinâmico de investigação e descoberta científica. Os autores acrescentam que duas funções distintas podem ser evidenciadas no processo de pesquisa: dados como materiais (*input*), que compreende a primeira parte

do processo, em que os dados são coletados e analisados, sendo provenientes de várias fontes, formas e formatos, constituindo um material para ser explorado e para levantar hipóteses. E os dados como resultado (*output*), que são os produzidos no decorrer do processo e no final, consequentemente estes dados são publicados como resultado da pesquisa.

O processo comumente chamado de ciclo de vida de dados compreende estágios por meio dos quais os dados se movem, da sua criação até a exclusão ou destruição (GOBEN; RASZEWSKI, 2015).

Ainda nesta temática existe a vertente que aborda os dados abertos, que seriam disponibilizados dentro de uma filosofia de ciência aberta para uso em outras pesquisas correlatas. Clobridge (2015) enfatiza que para o compartilhamento de dados em condições de ser potencialmente detectáveis e utilizáveis por outros, são necessárias boas práticas de pesquisa de gestão de dados. Tais ações envolvem passos do ciclo de vida dos dados, sendo necessário pensar sobre o processo inteiro de como eles são criados desde o início até o final. No mesmo pensamento, Darch et al. (2015) complementam que o primeiro passo para promover o compartilhamento de dados é promover práticas eficazes de gestão de dados a cada estágio do seu ciclo de vida.

Para Clobridge (2015), as fases do ciclo de vida dos dados estão relacionadas com coletar, descrever, organizar, arquivar, preservar, divulgar (para dados abertos). Ainda segundo a autora, essas fases nem sempre ocorrem na mesma ordem no decorrer do processo de pesquisa, pois os dados são dinâmicos, ocorrem à limpeza, codificação, reorganização, adicionamento, descarte e análises. Essas etapas podem acontecer ao mesmo tempo durante um período considerável, resultando em diversos arquivos complexos, por isso a necessidade de um gerenciamento adequado.

De certa forma, a gestão de dados de pesquisa elucida os preceitos de seleção das agências de fomento e outras estruturas, com a exigência de planos de gestão de dados, garantias de preservação em longo prazo e compartilhamento em acesso aberto (SCHÖPFEL et al., 2016).

De acordo com Hua et al. (2015), mesmo que a temática do gerenciamento e compartilhamento de dados de pesquisa científica seja um assunto debatido a nível mundial, grande parte dos pesquisadores não estão familiarizados em lidar de forma adequada com seus conjuntos de dados de pesquisa. Assim sendo, algumas bibliotecas oferecem treinamento para que os pesquisadores enfrentem os desafios encontrados durante o processo de pesquisa. Os autores exemplificam que na Biblioteca da Universidade de Cambridge e na *Australian National University Library*, o treinamento em gestão de dados está presente na alfabetização informacional.

Diante do diálogo extraído dos autores supramencionados, ficou evidenciado que os dados são para a pesquisa, por isso a necessidade de oferecer um tratamento adequado aos mesmos, de forma que possam ser úteis para novos estudos, além do mais, ficou claro que um ciclo de vida de dados compreende uma ferramenta metodológica essencial para a gestão de dados, ao possibilitar que todas as fases de um projeto de pesquisa possam ser acompanhadas.

2.1 MODELOS DE CICLO DE VIDA DOS DADOS

Um ciclo de vida de dados na visão de Rice e Southall (2016) mapeia a atividade de um pesquisador durante um projeto de pesquisa, de modo similar, ele delineia o caminho dos dados ou as ações necessárias a respeito dos mesmos, para possibilitar que a pesquisa avance para o próximo estágio.

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).* Assim, optamos por representar graficamente apenas estes dois modelos.

O DCC é um modelo de curadoria digital, que foi criado no Reino Unido, voltado para uma curadoria e preservação de dados bem sucedida. Os dados estão no centro do *Curation Lifecycle.* As ações estão divididas em três tipos: ações para todo o ciclo de vida, ações sequenciais e ações ocasionais (DCC, 2018). O que está expresso na Figura 1 a seguir.

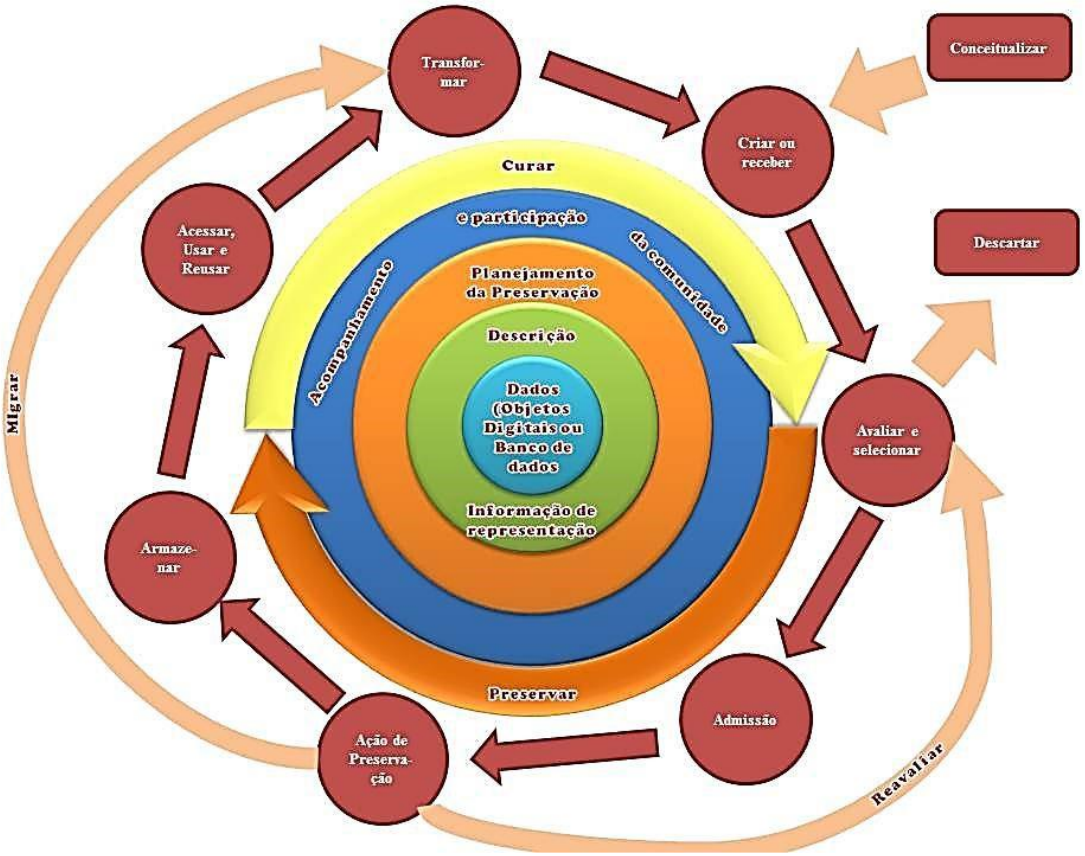
Os elementos pertencentes às ações para todo ciclo de vida compreendem as ações de: descrição e representação da informação, planejamento da preservação, vigilância e participação da comunidade, organizar e preservar. De acordo com Sayão e Sales (2012), as ações para todo o ciclo

de vida são nomeadas desta forma, porque permeiam todo o ciclo de vida da curadoria digital. Elas constituem atividades de planejamento que estão presentes de forma contínua.

As ações sequenciais envolvem os estágios de conceituar, criar ou receber, avaliar e selecionar, admissão, ação de preservação, armazenar, acessar usar e reusar e transformar. Para os autores Sayão e Sales (2012), tais ações necessitam ser realizadas repetidamente para garantir que o dado seja curado por meio de boas práticas. Por fim, as ações ocasionais, que para os autores fornecem estágios que são aplicados eventualmente, por meio das ações de eliminação, reavaliação e migração.

O modelo do *DataOne* proporciona gerenciamento e preservação de dados para uso e reuso, voltado para as ciências ambientais. Trata-se de uma iniciativa americana, que é financiada pelo *National Science Foundation* (NSF) (*DATAONE,* 2018).

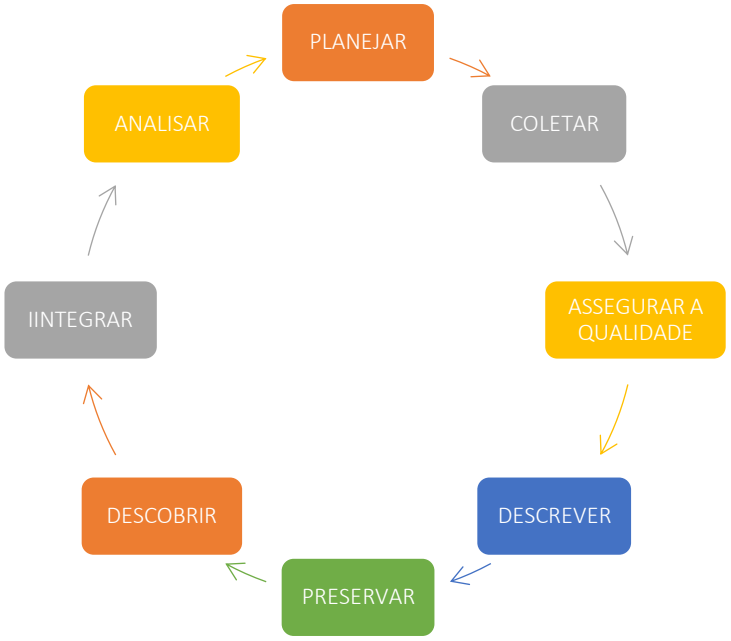
Figura 1 - Ciclo de vida da curadoria digital



Fonte: YAMAOKA, 2012

O modelo de ciclo de vida dos dados do *DataOne* envolve oito etapas, a primeira é **planejar**, neste momento se define quais os dados que serão gerados e como serão coletados e analisados. A etapa seguinte é **coletar**, as observações são trabalhadas de forma manual, por meio de sensores ou através de outros instrumentos, em que os dados são colocados em formato digital. A terceira etapa consiste em **assegurar a qualidade**, ocorre à garantia e o controle da qualidade dos dados, por meio de cheques e inspeções. A quarta etapa é **descrever**, os dados são documentados, descritos com a utilização de metadados adequados. Após este momento que é a chave para a compreensão futura dos dados, a quinta etapa é **preservar**, os dados são colocados em um arquivo adequado (centro de dados), pertinente a área de pesquisa, a próxima etapa é **descobrir**, os dados úteis potenciais são localizados e obtidos juntamente com as informações importantes a respeito dos mesmos, a sétima etapa é **integrar**, acontece a combinação dos dados de fontes distintas para a formação de um conjunto homogêneo de dados que pode ser analisado e por fim a última etapa é **analisar**, neste momento os dados são analisados (STRASSER et al., 2012). O que está evidenciado na Figura 2.

Figura 2: Ciclo de Vida dos Dados – *DataONE*



Fonte: Adaptado *DataOne* (2018)

Vardigan (2013) apresenta outro modelo, o *Data Documentation Initiative* (DDI), segundo a autora é uma iniciativa padrão pensada para documentar os dados no contexto social e nas ciências comportamentais. Hoyle et al. (2015) enfatizam que trata-se de uma inciativa desenvolvida em 1995, pela *Inter-university Consortium for Political and Social Research* (ICPSR), nos Estados Unidos, a qual também contou com o apoio da NSF. Uma das suas linhas de desenvolvimento é o DDI *lifecycle*, com um escopo amplo, envolve o ciclo de vida dos dados de pesquisa para a conceituação da coleção e processamento para a publicação de dados, além de outras funções. Os autores revelam que mesmo com as origens nas ciências sociais baseadas em pesquisas, o modelo DDI também pode ser usado por pesquisadores em outras áreas. De acordo com Sant'Ana (2013), este modelo contém 8 fases sequências, são elas: projeto, coleta, processamento, armazenamento, distribuição, recuperação, análise e reuso. Ele foi adotado pelo sistema de bibliotecas do *Massachusetts Institute of Technology* – MIT e fornece um modelo de referência para criação de outros modelos.

Diante do uso e acesso intenso de dados, no olhar de Sant'Ana (2016), a Ciência da Informação pode e deve contribuir no processo de otimização e uso dos dados, para tanto, o autor desenvolveu no cenário brasileiro, um modelo de ciclo de vida de dados voltado para a CI, que contempla as fases de coleta, armazenamento, recuperação e descarte e os fatores que estão presentes em cada uma delas, os quais são: privacidade, integração, qualidade, direitos autorais, disseminação e preservação.

Além dos modelos do *DataOne*, DCC, DDI e Santa'Ana (2013), podemos citar o modelo da Jisc[1] (Joint Information Systems Committee) (2016), que é uma organização de membros, que fornece soluções digitais para a educação e pesquisa no Reino Unido, trabalha com um diagrama de ciclo de vida de dados de pesquisa voltado para o seu gerenciamento. Rice e Southall (2016) ao referenciar o modelo em questão, revelam que uma vantagem de um diagrama circular é apresentar como os dados continuam a existir após um projeto de pesquisa, embora que muitas vezes seja no projeto de outro pesquisador.

[1] https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data

Podemos ainda citar outro modelo do *Imperial College London Library RDM Workflow*[2], cujo autor é Barnes (2016), também evidenciado por Rice e Southall (2016), os quais revelam que ainda existem diversos modelos de ciclo de vida de dados.

Desta forma, os modelos supramencionados mesmo com suas particularidades e objetivos, são iniciativas voltadas para apoiar o desafio de gerenciar efetivamente os dados de pesquisas, de forma que os elementos pertencentes aos ciclos de vida dos dados possam ser acompanhados em todas as fases necessárias para um projeto de acordo com a necessidade dos pesquisadores.

3 PROCEDIMENTOS METODOLÓGICOS

O presente estudo trata-se de uma abordagem quantitativa, pois de forma objetiva foi possível analisar a relação entre as varáveis, que podem ser mensuradas por instrumentos, possibilitando que os dados possam ser examinados através de procedimentos estatísticos (RICHARDSON, 2017).

A realização desta pesquisa teve como fonte de dados os metadados dos artigos encontrados nas bases de dados: Emerald, LISA e LISTA. A coleta do estudo em questão foi realizada no 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 pesquisado sobre a temática nos últimos 5 (cinco) anos. Por fim, os tipos de documentos compreenderam os artigos de arquivos abertos.

Após o *download* dos arquivos de dados (formato pdf) foi construída uma tabela com a quantidade de artigos encontrados nos anos previstos na amostra, provenientes de cada base de dados, o que está ilustrado na Tabela 1 a seguir:

Tabela 1: Levantamento nas bases de dados sobre a quantidade de artigos disponibilizados contendo o termo "*Data life cycle"* no período de 2013-2018.

Nome da Base	2013	2014	2015	2016	2017	2018	TOTAL
Emerald	-	3	8	4	4	3	22
Lisa	-	2	4	4	3	1	14
Lista	2	5	14	11	-	3	35
TOTAL	2	10	26	19	7	7	71

Fonte: Dados da pesquisa (2018)

Os dados revelam que a base que mais indexou no período em questão foi a LISTA, pois através dela foi possível recuperar 35 artigos. Em 2015 houve um avanço nas produções, o que pode ser verificado nas três bases. Em 2016 o volume foi significativo. Apresentou uma queda em 2017 na base LISTA, uma vez que, não houve indexações. Em 2018 as referidas bases continuaram disponibilizando artigos sobre a temática. Desta forma, o tema é alvo de discussão no cenário atual.

Os artigos oriundos da busca bibliográfica realizada na Emerald, LISA, LISTA foram extraídos para a ferramenta ZOTERO[3], a qual é um *software* livre, de fácil acesso, que ajuda na coleta, organização, citação e compartilhamento. Com uso desta ferramenta 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.

As bases citadas foram selecionadas, porque ambas têm a Ciência da Informação como área de concentração, contudo também indexam conteúdo de outras áreas. De acordo com o portal de periódicos da Capes (2018), a Emerald disponibiliza coleção de publicações periódicas, a LISTA indexa mais de 500 periódicos científicos, incluindo texto completo de mais de 240 periódicos científicos e a LISA é uma base de dados internacional que indexa mais de 400 títulos de periódicos provenientes de

[2] https://zenodo.org/record/54000#.W5MbI0ZKjIU
[3] https://www.zotero.org/

mais de 68 países e em mais 20 idiomas distintos. O que mostra que são bases relevantes para a área da Ciência da Informação.

Desta forma, trata-se de um estudo bibliográfico, pois Segundo Gil (2017), neste tipo de pesquisa é realizado um aporte em material já publicado e com os novos formatos que possibilitam a disseminação da informação, o material disponibilizado na internet passou a ser uma fonte de tal modalidade. Trata-se ainda de uma pesquisa exploratória que buscou trazer clareza de um tema pouco explorado (GIL, 2019).

Na construção do aporte teórico, foram priorizadas as referências oriundas da pesquisa dos autores que publicaram mais de uma vez sobre a temática. Pela insuficiência de citações, outros textos da investigação dos autores com apenas uma produção também foram usados como referência, além de outros encontrados na literatura. Cabe salientar que parte dos artigos selecionados para o referencial teórico trata de dados científicos, foram objetos recuperados já neste trabalho.

Para uma análise mais refinada, os textos foram codificados para extrair as citações que tratavam sobre o *Data life cycle*, e os artigos dos autores que mais publicaram foram enviados para o *software* QDA Miner[4]**, o qual é** um pacote de *software* de análise de dados qualitativo, podendo ser usado para codificar, anotar, recuperar e analisar pequenas e grandes coleções de documentos e imagens.

No momento seguinte todos os textos recuperados na pesquisa foram enviados para a ferramenta, em que foi possível fazer o levantamento das formas variantes do termo encontradas nos textos, assim como apresentar as palavras referentes à temática que mais se destacaram.

4 DISCUSSÕES

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% 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 tabela 2 a seguir destaca os autores que publicaram mais de um artigo.

Tabela 2 Autores que mais publicaram com o termo "*Data life cycle*".

AUTORES		ARTIGOS
Número	**Nome**	**PUBLICADOS**
1	Vardigan, Mary	4
2	Borgman, Christine L.	2
3	Češarek, Ana	2
4	Darch, Peter T.	2
5	Harris, Sasekea	2
6	Hua, Xiaoqin	2
7	Ionescu, Sanda	2
8	Južnič, Primož	2
9	Koler-Povh, Teja	2
10	Li, Xin	2
11	Malleret, Cécile	2
12	Prost, Hélène	2
13	Sands, Ashley E.	2
14	Schöpfel, Joachim	2
15	Si, Li	2
16	Traweek, Sharon	2
17	Wallis1, Jillian C.	2
18	Xing, Wenming	2
19	Zhuang, Xiaozhe	2

Fonte: Dados da pesquisa (2018)

Os dados da tabela 2 mostram que a autora que se destacou na quantidade de artigos sobre o tema foi Mary Vardigan, produziu 4 (quatro) artigos nos períodos de 2013 a 2016, no periódico

[4] https://provalisresearch.com/products/qualitative-data-analysis-software/

americano IASSIST Quarterly, sendo que a produção de 2013 foi apenas de sua autoria e as demais envolveram outros pesquisadores, sendo duas delas com coautoria de Ionescu.

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 em coautoria. 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 2 (dois) artigos e Li que desenvolveu suas pesquisas com outros pesquisadores que não estão citados na Tabela 2.

A análise possibilitou levantar os periódicos que apresentaram mais produções no período sobre a temática, foram listados na tabela 3 somente os que têm acima de uma publicação.

Tabela 3: Distribuição dos periódicos por países, no período de 2013-2018.

TÍTULO DE PUBLICAÇÃO	QUANTIDADE	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 produções científicas ocorreu no periódico americano IASSIST[5] Quarterly, ao apresentar 8 produções, o que está alinhado Vardigan que teve o maior número de publicações no referido periódico.

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 3 a seguir.

Diante dos dados apresentados, é possível identificar que 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 perfeiçoar o acesso à informação científica, no que tange aos artigos de pesquisa científica e aos dados de pesquisa. Trata-se de uma estratégia europeia que tem o conhecimento e a inovação como impulsionadores do crescimento econômico. Em um contexto digital de *Open Acess* (HORIZON, 2018).

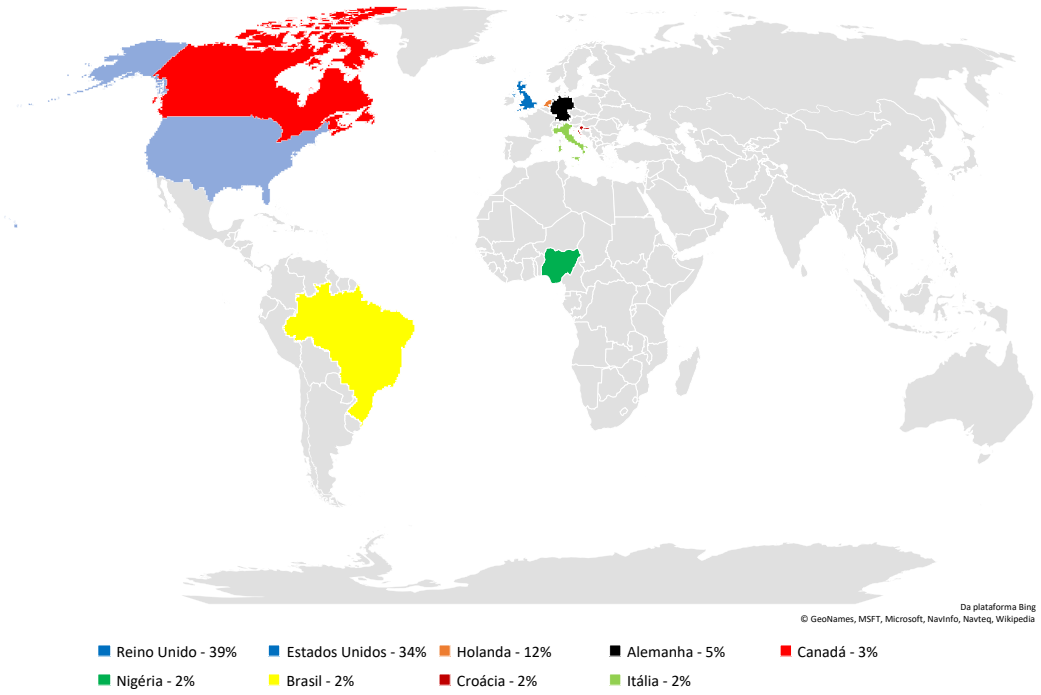
Nesse cenário, é possível observar que o Reino Unido se destacou com um percentual de 39%, o que pode remeter a ser um local de destaque na criação de ciclos de vida de dados, como por exemplo, o DCC, que segundo Goben e Raszewski (2015) é um dos modelos mais populares. A iniciativa da Jisc e o do *Imperial College London Library RDM Workflow* são exemplos de modelos desenvolvidos no Reino Unido, os quais foram evidenciados por Rice e Southall (2016). 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 a treinamentos para os pesquisadores gerenciarem seus dados.

Posteriormente aparece o continente americano com 39% das produções, sendo 34% dos Estados Unidos. Diante desta realidade, é possível constatar a presença de iniciativas que visam a gestão

[5] International Association for Social Science Information Service and Technology

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 (2013). Por fim, foi possível identificar um percentual de 2% no continente africano, especificamente na Nigéria.

Figura 3: Países que publicaram sobre *data life cycle* no período de 2013-2018.

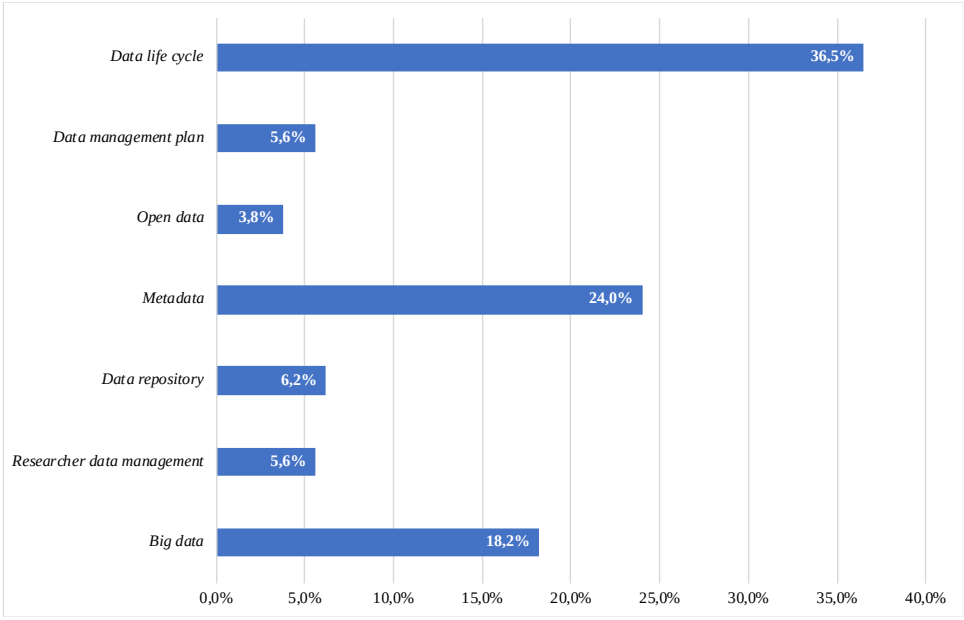


Dados da pesquisa (2018)

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

Tomando como base as palavras-chave dos artigos indicadas pelos autores que publicaram mais de um artigo, por meio da ferramenta QDA MINER, foi possível efetuar a 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 manegament plan* (plano de gestão de dados), para verificar a frequência desses termos em todos os arquivos recuperados, nos títulos, palavras-chave e corpo do texto, 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 QDAMiner.

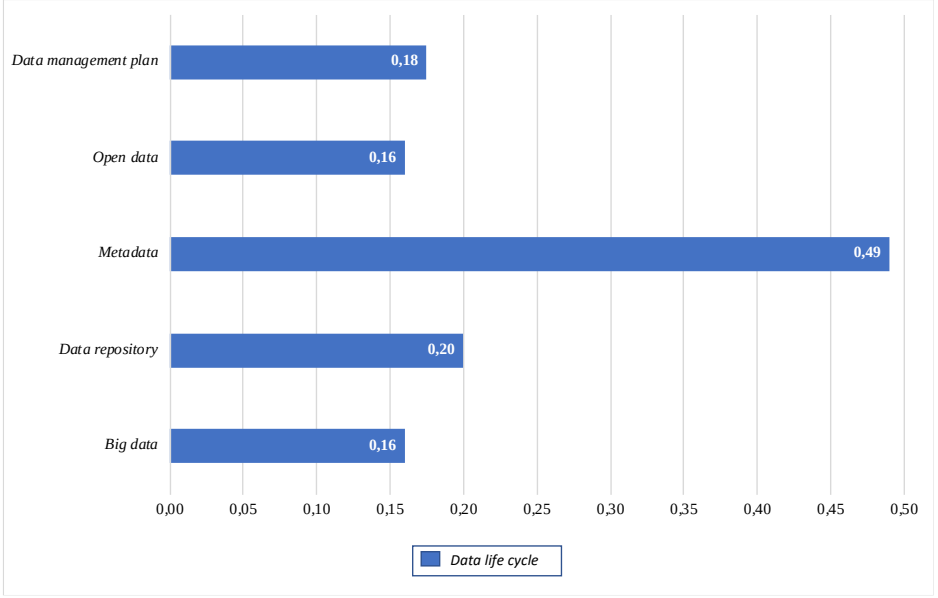


Fonte: Dados da pesquisa (2018)

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-chaves, o que possibilitou a extração de segmentos relevantes contidos nos textos para o apoio teórico da pesquisa.

O Gráfico 2 a seguir, apresenta a ocorrência 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 um dos termos mais citados, na sequencia do *data life cycle* e do *metadata* 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.

Foi realizada uma busca pelo do termo "*data life cycle*". A partir de uma investigação refinada com a eliminação de títulos, referências e partes dos textos não úteis à pesquisa, foi possível identificar 68 ocorrências entre textos e figuras encontrados em 8 artigos, o que representa 61,5% dos casos pesquisados. Assim, a ferramenta possibilitou recuperar os segmentos selecionados e identificar os artigos em que se encontravam. Isso contribuiu com a construção da fundamentação teórica, em que mais de 50% dos autores que se encaixaram no patamar dos que publicaram mais de uma vez, foram citados, embora parte deles não apareça no texto, pois alguns se encontram nas citações de mais de três autores com a utilização do et al.

A partir da busca pelo termo supracitado nos artigos recuperados, foi possível identificar três variantes do mesmo: *data life-cycle*, *data lifecycle* e *big data life cycle*. Assim sendo, existe mais de uma forma de escrevê-lo na literatura internacional. Segundo a versão *online* do dicionário de *Cambridge*, o termo *life cycle* (ciclo de vida) significa uma série de mudanças pelas quais algo passa durante a sua existência. Analogamente, os ciclos de vida de dados - evidenciados no referido estudo – representam os estágios de evolução pelos quais os dados de uma pesquisa científica percorrem no decorrer de sua história.

5 CONSIDERAÇÕES FINAIS

O presente trabalho contribuiu para atingir o objetivo da pesquisa que foi realizar uma revisão bibliográfica sobre o termo *Data life cycle,* na literatura concentrada na área da Ciência da Informação ao facilitar o entendimento da área pelos pesquisadores que investigam o tema. Evidenciou-se que temática ainda é incipiente na área da Ciência da Informação, uma vez que, a quantidade de produções não foi expressiva no período analisado. O termo é discutido no cenário internacional, com destaque para o Reino Unido com um maior de número de produções e Mary Vardigan foi quem mais produziu no período estudado.

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.

Portanto, a ciência da informação pode ser uma referência para outras ciências ao buscar explorar mais sobre o ciclo de vida dos dados, haja vista a importância do tema evidenciado no cenário internacional.

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, assim como a origem dos países que os trabalhos foram produzidos e não apenas publicados. Sugerimos ainda pesquisas que por meio de uma revisão sistemática da literatura apresentem de forma crítica as contribuições expressas na literatura sobre a temática investigada.

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 ago. 2018.

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

CAMBRIDGE DICTIONARY. Disponível em: https://dictionary.cambridge.org/pt/dicionario/ingles/life-cycle. Acesso em: 19 set. 2018.

COYNE, E. M.; COYNE, J. G.; WALKER, K. B. Big Data information governance by accountants. **International Journal of Accounting & Information Management**, v. 26, n. 1, p. 153–170, 8 fev. 2018. Disponível em: https://www-emeraldinsight-com.ez15.periodicos.capes.gov.br/doi/full/10.1108/IJAIM-01-2017-0006. Acesso em: 20 ago. 2018.
CLOBRIDGE, A. Open Data: Shining a Light on Data Management Practices. **Online Searcher**, v. 39, n. 4, p. 68–70, 7 ago. 2015. Disponível em: https://search-proquest.ez15.periodicos.capes.gov.br/docview/1702929531?pq-origsite=primo. Acesso em: 20 ago. 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.
DARCH, P. et al. What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and beyond. **International Journal on Digital Libraries**, v. 16, n. 1, p. 61–77, mai. 2015. Disponível em: https://link-springer-com.ez15.periodicos.capes.gov.br/article/10.1007/s00799-015-0137-3. Acesso em 20 ago. 2018.
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.
DATAONE. **What is DataONE?** Online. Disponível em: https://www.dataone.org/what-dataone. 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.
FEDERER, L. Research data management in the age of big data: roles and opportunities for librarians. **Information Services & Use**, v. 36, n. 1/2, p. 35–43, jan. 2016. DOI 10.3233/ISU-160797
GIL, A. C. **Como elaborar projetos de pesquisa**. 6. ed. São Paulo: Atlas, 2017.
GIL, A.C. **Métodos e técnicas de pesquisa social**. 7 ed. São Paulo: Atlas, 2019.
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. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4279933/. Acesso em 17 de jul. 2018. Doi: http://dx.doi.org/10.3163/1536-5050.103.1.008.
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.
HOYLE, L. *et al*. DDI and Enhanced Data Citation. **IASSIST Quarterly**, v. 39, n. 3, p. 30–46, 2015.
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. Disponível em: https://www-emeraldinsight-com.ez15.periodicos.capes.gov.br/doi/full/10.1108/EL-07-2013-0130. Acesso em: 20 ago. 2018.
MAYER-SCHÖNBERGER, V.; CUKIER, K.. **Big data**: A revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt, 2013.
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.
RICHARDSON, R.J. **Pesquisa social:** métodos e técnicas. 4. ed. São Paulo: Atlas, 2017.
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. Disponível: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940. Acesso em: 20 jun. 2018.
SAYÃO, L. F. SALES, L. F. Curadoria geral: um novo patamar para a preservação de dados digitais de pesquisa. **Informação & Sociedade**. v. 22, n. 3. João Pessoa. Set./Dez. 2012. p. 179-191. Disponível em: http://www.ies.ufpb.br/ojs/index.php/ies/article/viewFile/12224/8586. Acesso em: 20 ago. 2018.
SCHÖPFEL, J. *et al*. Dissertations and Data. **Grey Journal (TGJ)**, v. 12, n. 3, p. 126–148, set. 2016.

STRASSER, C. et al. **Primer on Data Management**: What you always wanted to know. California: CDL, 2012. Disponível em: http://escholarship.org/uc/item/7tf5q7n3#page-1. Acesso em: 13 jun. 2018.
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.