XX ENANCIB
21 a 25 Outubro/2019 – Florianópolis
A Ciência da Informação e a era da Ciência de Dados

ISSN 2177-3688

GT-8 – INFORMAÇÃO E TECNOLOGIA

CICLO DE VIDA DE DADOS NO PROCESSO DE PUBLICAÇÃO E ACESSO À PRODUÇÃO CIENTÍFICA

DATA LIFE CYCLE IN THE PUBLICATION PROCESS AND ACCESS TO SCIENTIFIC PRODUCTION

Emanuelle Torino - Universidade Tecnológica Federal do Paraná; Unesp
Silvana Aparecida Borsetti Gregorio Vidotti[1] - Universidade Estadual Paulista (Unesp)
Ricardo César Gonçalves Sant'Ana - Universidade Estadual Paulista (Unesp)

Modalidade: Trabalho Completo

Resumo: O advento das tecnologias de informação e comunicação amplia a disponibilização de ambientes digitais para produção, armazenamento e publicação de dados e informações, o que torna necessário alterar as formas de análise, coleta, processamento, armazenamento, disseminação, recuperação e acesso, visando otimizar processos e beneficiar os atores envolvidos, quer sejam produtores ou consumidores. O presente estudo objetiva apresentar os ciclos de vida de dados e atores envolvidos no processo de publicação e acesso à produção científica, bem como identificar seu principal ator e fornecer a ele contribuições da Ciência da Informação. Para tanto, foi realizada a pesquisa bibliográfica para embasamento teórico-conceitual do estudo e a pesquisa descritiva para explicitar os ciclos de vida de dados estudados. Como resultado apresentam-se as formas de coleta, armazenamento e recuperação de dados que compõem o campo informacional descrito, de forma a fornecer subsídios para o reuso de dados dos ciclos apresentados em diferentes contextos, bem como a identificação do principal ator, os fatores que o influenciam e são influenciados por ele, além de fornecer contribuições da Ciência da Informação.

Palavras-Chave: Ciclo de vida dos dados; Publicação científica; Reuso de dados.

[1] Bolsista produtividade em pesquisa CNPq (PQ2).

Abstract: The advent of information and communication technologies expands the availability of digital environments for the production, storage and publication of data and information, which makes it necessary to change the forms of analysis, collection, processing, storage, dissemination, retrieval and access, aiming to optimize processes and benefit the actors involved, whether they are producers or consumers. This study aims to present the life cycles of data and actors involved in the process of publication and access to scientific production, as well as identify their main actor and make contributions from Information Science. To this end, a bibliographic research was conducted for the theoretical and conceptual basis of the study and a descriptive research to explain the life cycles of the data studied. As a result we present the forms of data collection, storage and retrieval that make up the described informational field, in order to provide subsidies for the reuse of data from the cycles presented in different contexts, as well as the identification of the main actor, the factors that influence and are influenced by it, and provide contributions from Information Science.

Keywords: Data life cycle; Scientific publication; Data Reuse.

1 INTRODUÇÃO

O processo de comunicação científica consiste na interação entre pesquisadores de diferentes áreas do conhecimento, utilizando-se de canais formais e informais (CÔRTES, 2006; GARVEY, 1979; KLING; CALLAHAN, 2003; MEADOWS, 1999), e compreende atividades concernentes à produção, disseminação e uso da informação, iniciada pelo processo de pesquisa e criação de conhecimento até a aceitação dos resultados pela comunidade científica (GARVEY, 1979).

Desde a Antiguidade buscam-se ferramentas para armazenamento e disponibilização de conhecimento, tornando-o acessível e permanente, fato esse que oportunizou que os periódicos científicos se constituíssem como a principal estrutura de comunicação científica (WEITZEL, 2006). A esses foi estabelecida ainda a prerrogativa de validação por pares (*peer review*), além da publicação ágil e, portanto atualizada, de resultados de pesquisa. O *input* desse processo é realizado pelo próprio pesquisador, que figura como autor e a quem compete a submissão do seu resultado de pesquisa à análise de um periódico, que atua como fonte publicação.

O reconhecimento da qualidade e rigor científico de tais publicações é estabelecido pelo aceite do periódico em indexadores relevantes na área de abrangência, que após criteriosa análise disponibilizam em sua base de dados os metadados das fontes publicadoras e que conferem a estas credibilidade e visibilidade e, permitem ao usuário a recuperação de informação proveniente de múltiplas fontes.

Nesse cenário, de forma complementar a esses ambientes informacionais digitais, são inseridos os repositórios digitais que, segundo sua abordagem e objetivo, reúnem e

armazenam objetos digitais, de diferentes tipologias, visando assegurar a ampliação de visibilidade e preservação em longo prazo (CAMARGO; VIDOTTI, 2011; LEITE, 2009; LYNCH, 2003).

O advento das tecnologias de informação e comunicação (TIC) amplia a disponibilização de ambientes digitais para produção, armazenamento e publicação de dados e informações, o que torna necessário alterar as formas de análise, coleta, processamento, armazenamento, disseminação, recuperação e acesso, visando otimizar processos e beneficiar os atores envolvidos, quer sejam produtores ou consumidores. Sant'Ana (2016) destaca que é necessário estabelecer pontes entre os sujeitos informacionais e suas necessidades para a efetividade da disponibilização de dados, sob pena de prejuízos danosos em detrimento aos benefícios que podem ser gerados pelo acesso.

Nesse contexto, o presente estudo objetiva apresentar os ciclos de vida de dados e atores envolvidos no processo de publicação e acesso à produção científica, bem como identificar seu principal ator e fornecer a ele contribuições da Ciência da Informação.

Para tanto, utiliza como base o ciclo de vida dos dados (CVD) proposto por Sant'Ana (2013; 2016) para a delimitação de fases de acesso e uso de dados, mantendo-os como centro do CVD. Busca enfatizar ainda, além dos dados, o principal ator envolvido no processo de publicação e acesso à produção científica, de forma a explicitar as contribuições da Ciência da Informação para esse campo informacional.

As fases do CVD preconizadas por Sant'Ana (2016) consistem em:

a) coleta: obtenção dos dados necessários para uma determinada demanda e que deve considerar aspectos como escopo, resultados esperados, fontes, formatos de dados, tratamento necessário à utilização, privacidade, possibilidade de integração com outros dados, integridade física e lógica, procedência, direitos autorais, possibilidade de identificação e recuperação futura, presença de dados que permitam a manipulação e acesso;
b) armazenamento: manutenção dos dados coletados para uso futuro. Para tanto é necessário conhecer os dados coletados, definir quais serão mantidos, qual estrutura será utilizada para armazenamento, assegurar a permanência dos dados coletados para fornecer contexto ao *dataset*, privacidade dos dados, qualidade dos

dados e aspectos que possibilitem a encontrabilidade dos *datasets*, fatores esses necessários à utilização em longo prazo;

c) recuperação: uma vez coletados e armazenados, os dados mantidos precisam ser recuperados para que possam ser acessados, interpretados e utilizados. Nesta fase é imprescindível definir quais dados armazenados serão recuperáveis, público-alvo, escopo, acesso à base de armazenamento ou estabelecimento de base específica para a recuperação, frequência de atualização, níveis de acesso, direitos de acesso, qualidade e privacidade dos dados, possibilidade de integração de *datasets* e recuperação para utilização em longo prazo;
d) descarte: as fases anteriores definirão os dados que não precisam ser mantidos e devem ser excluídos da base, ação que consiste na limpeza ou desativação dos dados. Para tanto, deve-se ter clareza de quais dados devem ser descartados, se foram persistidos, se foram duplicados em outras bases, como assegurar o descarte considerando a possibilidade de ocultá-los, impactos do descarte em outros dados a eles ligados e aos disponíveis na própria base, além dos impactos na recuperação e no acesso.

Destaca ainda o autor que todas as fases estão permeadas pelos fatores: privacidade, integração, qualidade, direitos autorais, disseminação e preservação (SANT'ANA, 2013; 2016).

Cada uma das fases requer conhecimentos e habilidades específicos, provenientes de áreas como a Ciência da Informação e a Ciência da Computação. Além destas fases, o processo é motivado, pelo que o Sant'Ana (2016) considera usuários, e, tem como atores:

a) detentor: pessoa ou organização que estruturou o ciclo de vida dos dados e que, portanto, possui todo o conhecimento das suas fases. Responsável pela manutenção do ambiente e quem toma decisões acerca das formas de coleta, armazenamento, recuperação e descarte; e que pode realizar a atividade ou delegá-la a um intermediário, que realizará todas ou algumas fases do processo;
b) intermediário: pessoa ou organização que desenvolveu a aplicação e possui conhecimentos acerca da tecnologia. Além desse, considera-se intermediário a pessoa ou organização que atua sobre o CVD por solicitação do detentor;

c) referenciado: pessoa, organização ou atividade vinculada ou mencionada nos dados a serem disponibilizados e, por isso, consideradas referenciadas;
d) usuário: sujeito cadastrado no sistema, que interage com o CVD, e que fornece seus dados consciente ou inconscientemente para armazenamento.

Para o desenvolvimento do estudo foi realizada pesquisa bibliográfica para embasamento teórico-conceitual e pesquisa descritiva para explicitar os ciclos de vida de dados no processo de publicação e acesso à produção científica, utilizando como ferramentas o Open Journal System (OJS) para periódicos, o DSpace para repositórios digitais, além das bases Scientific Electronic Library Online (SciELO) e Scopus como indexadores de periódicos.

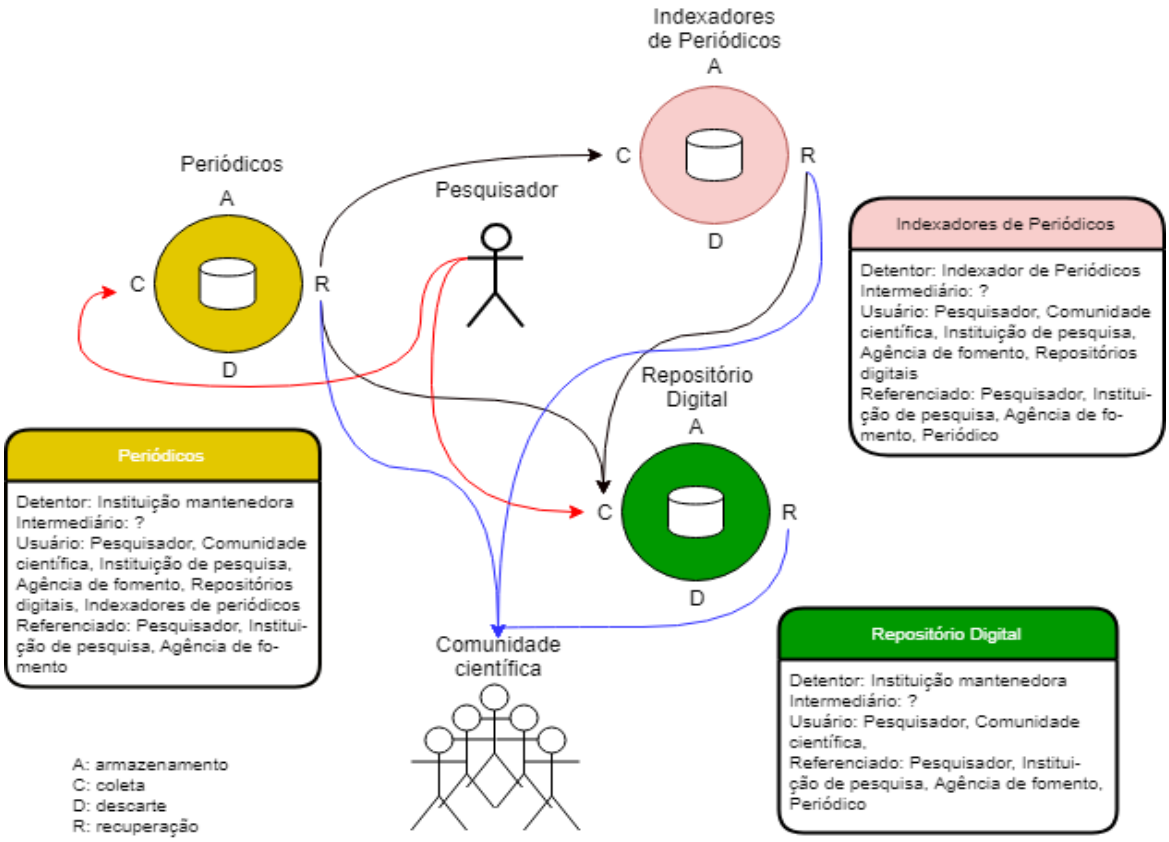
No que tange à pesquisa descritiva, foram analisadas as informações disponíveis na documentação dos sistemas, suas páginas web e o próprio uso. Tornou-se possível identificar o principal ator no campo informacional apresentado e elencar os fatores que o impactam diretamente, de forma a discutir seu papel e apresentar as contribuições da Ciência da Informação.

2 ANÁLISE DOS CICLOS DE VIDA DE DADOS

Considerando-se os aspectos anteriormente abordados, o presente estudo concentra-se no campo informacional relativo ao processo de publicação e acesso à produção científica, utilizando-se de fontes de publicação (periódicos) e ampliação de visibilidade (indexadores de periódicos e repositórios digitais), tendo como atores os pesquisadores e a comunidade científica, conforme explicitado na Figura 1. Nesse campo informacional, o principal ator identificado é o pesquisador, que inicia o processo com a submissão de um manuscrito a um periódico.

Na Figura 1 as setas representadas por diferentes tonalidades indicam as formas de relacionamento entre os CVDs e seus atores no campo informacional. As setas vermelhas indicam que o pesquisador pode alimentar a coleta dos CVDs dos periódicos e dos repositórios digitais, cuja representação do objeto digital é realizada manualmente e, deve atender aos objetivos e escopo do ciclo determinado.

Figura 1: Ciclos de Vida de Dados e Atores no Processo de Publicação e Acesso à Produção Científica



Fonte: Autoria própria (2019).

As setas pretas indicam que a fase de recuperação dos CVDs pode alimentar a coleta de outros ciclos, atividade esta que pode ser realizada por uma aplicação computacional que favorece o intercâmbio de registros e objetos entre diferentes fontes, ação definida por VIDOTTI *et al.* (2016) como coleta automática. Já as setas azuis indicam que os três CVDs representados na Figura 1 possibilitam a recuperação da informação por humanos, por meio da interface, utilizando elementos de navegação e/ou mecanismo de busca.

2.1 Ciclo de Vida de Dados em Periódicos Científicos

Os periódicos constituem-se na principal fonte de comunicação de resultados de pesquisa e, tiveram significativo advento com a evolução tecnológica, sobretudo em ambientes digitais, o que tornou mais ágil e eficiente o processo e efetivo o seu ciclo de vida.

O funcionamento é definido pela política editorial, que individualiza as práticas em cada publicação. Contudo, algumas delas podem ser generalizadas, a exemplo da coleta e da recuperação.

Para este estudo, toma-se como base o Open Journal System (OJS), software mais utilizado para a gestão e publicação de periódicos científicos no Brasil. O OJS é uma ferramenta desenvolvida e distribuída pelo Public Knowledge Project (PKP) vinculado à University of British Columbia, no Canadá. Desenvolvido em linguagem PHP, utilizando banco de dados MySQL ou PostgreSQL, o software é distribuído utilizando uma licença GNU General Public License (GLP) e sua documentação uma licença Creative Commons, constituindo-se em uma iniciativa de código aberto para que seja melhorado pela comunidade, cujas contribuições podem ser compartilhadas (PUBLIC KNOWLEDGE PROJECT, 2018a; 2018b).

Ao considerarmos o CVD de periódicos gerenciados pelo OJS, o fluxo é iniciado pela coleta, alimentada pelo 'pesquisador' que, atendidas as políticas editoriais, submete o manuscrito (objeto digital), que pode ser acompanhado de material suplementar, para análise. Nesse processo, utiliza metadados no padrão Dublin Core para a representação da informação contida no objeto digital, bem como dos autores (pesquisadores). Destaca-se que a instalação *default* do OJS disponibiliza uma estrutura padrão de metadados para a representação, que pode ser alterada a critério do detentor.

A partir do preenchimento dos metadados, o pesquisador atribui contexto semântico ao objeto disponibilizado, favorecendo o processo de armazenamento, interpretação, recuperação, encontrabilidade e ligação dos dados.

Nesse contexto é relevante mencionar que a disponibilização de periódicos em ambiente web requer tratamento específico para que os dados possam atender às boas práticas para a produção e disponibilização de dados na web, preconizadas pelo World Wide Web Consortium (W3C) (LÓSCIO; BURLE; CALEGARI, 2017).

O OJS não limita o formato do objeto digital submetido, contudo, em geral, os periódicos solicitam que seja um arquivo editável (.doc, .docx, .odt), e, disponibilizam para a recuperação arquivos portáveis (.pdf).

Visando o atendimento à questão da avaliação por pares (*peer review*), tradicionalmente efetuada às cegas, o OJS altera automaticamente o nome do arquivo por meio da correspondência às etapas do fluxo editorial, estando sempre iniciado pelo número de identificação (ID) da submissão no sistema. O mesmo arquivo é automaticamente

renomeado durante processo de avaliação, à medida que avança no fluxo editorial, notadamente constituído por submissão, designação a editor, designação a avaliador, versão do autor, versão do editor, revisão de prova, leiaute, até a composição final da versão a ser publicada e que estará disponível para acesso. Nessas etapas, o arquivo mantém na sua nomenclatura o ID da submissão e altera a sigla, de acordo com a etapa no fluxo editorial, utilizando-se SM, RV, ED, CE, PB, alusivas respectivamente à: submissão, revisão, edição, leitura de prova e publicado, caracterizando o versionamento.

Por *default* no OJS todas as versões dos objetos digitais, bem como as decisões tomadas ao longo do fluxo, desde a submissão até a publicação são armazenados na base de dados. Desta forma, compete ao detentor a definição sobre o descarte dos dados.

A base de dados do OJS pode ser MySQL ou PostgreSQL, cabendo à instituição mantenedora (detentor) a definição, de acordo com a infraestrutura disponível. Os dados armazenados podem ser exibidos pelo protocolo Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), que possibilita a recuperação da informação em pacotes por aplicações computacionais, em diferentes padrões de metadados, como Dublin Core (DC), National Library of Medicine (NLM) e Machine-Readable Cataloging (MARC). Permite ainda a coleta por motores de busca, a exemplo do Google e Google Acadêmico. De igual maneira, o sistema disponibiliza interface na qual a recuperação da informação por humanos ocorre por meio de elementos de navegação e/ou do mecanismo de busca.

No CVD de periódicos, o detentor é a instituição mantenedora, que determina o processo de negócios, bem como as formas de coleta, armazenamento, descarte e recuperação, que poderão ser operacionalizadas por ele ou pelo intermediário, que pode ser parte da infraestrutura ou uma estrutura parceira.

No caso do OJS, em primeira instância, o intermediário é o PKP, enquanto desenvolvedor do sistema, podendo esse papel ser compartilhado com outros atores que tenham realizado alterações no código fonte do software.

Os referenciados são o(s) pesquisador(es) que assumem o papel de autor(es) do(s) trabalho(s) publicado(s), sua(s) instituição(ões) de afiliação e a(s) agência(s) de fomento que apoiou(aram) o desenvolvimento da pesquisa.

Os usuários podem ser humanos ou agentes computacionais, que se utilizam da fase de recuperação, realizada individualmente ou em lotes, atendendo ao processamento

humano e de máquina, sendo que esse último pode alimentar a coleta de outros CVDs, conforme descreveremos a seguir.

No que tange à segurança, o acesso dos usuários no sistema é configurado de acordo com as funções exercidas no processo, considerando que um usuário pode assumir diferentes papéis (como: autor, avaliador, editor, leitor) no mesmo periódico, bem como funções idênticas ou diferentes em cada uma das publicações que podem ser geridas pelo OJS. Na documentação do sistema há um guia de proteção de dados disponível para o usuário[2].

2.2 Ciclo de Vida de Dados em Indexadores de Periódicos Científicos

Os indexadores de periódicos cujo surgimento esteve atrelado ao expressivo aumento das publicações periódicas impressas, atuam como uma espécie de catálogo para a recuperação dos artigos publicados. As alterações tecnológicas modificaram o suporte dos periódicos para ambientes digitais e, com isso, os indexadores de periódicos, passaram a atuar como provedores de serviços e/ou de dados, que, a partir de criteriosa análise embasada em políticas próprias, chancelam a qualidade de periódicos científicos e, com isso, ampliam sua visibilidade e impacto.

Tais fontes são majoritariamente mantidas por potências editoriais que oferecem à publicação indexada, além da sua chancela, uma plataforma de acesso para os usuários, com aspectos tecnológicos benéficos a esses atores, a exemplo de múltiplas fontes de dados, sistemas de recomendação, área para armazenamento de histórico de buscas, entre outros.

Considerando que tais fontes são, em geral privadas, não é possível explicitar algumas etapas do seu CVD, embora seja possível afirmar que os dados e objetos são coletados dos periódicos utilizando Application Programming Interface (APIs) específicas e armazenados em bancos de dados para a recuperação. Tal ação pode ser realizada como um processo de coleta constante dos dados, após a publicação nos periódicos, ou ainda mediante envio de lotes de dados por esses gerados, considerando os requisitos especificados pelos detentores desse CVD.

Como limitação de análise do presente estudo, não é possível afirmar como os dados são armazenados, recuperados e se são descartados. Nesse sentido, é sabido que algumas

[2] Disponível em: https://docs.pkp.sfu.ca/gdpr/en/. Acesso em: 01 nov. 2018.

publicações são aceitas por um período temporal, porém, não há informações de como os dados são descartados ou ocultados quando deixam de ser parte dos conteúdos indexados.

Os detentores, a exemplo do CVD de periódicos, são as instituições mantenedoras, que possuem a prerrogativa de delegar a administração aos intermediários, que desenvolvem a aplicação e/ou atuam em uma ou mais fases do CVD. Os referenciados são o(s) pesquisador(es) que assumem o papel de autor(es) do(s) trabalho(s) publicado(s), sua(s) instituição(ões) de afiliação, a(s) agência(s) de fomento que apoiou(aram) o desenvolvimento da pesquisa e o periódico que publicou a contribuição. Os usuários podem ser humanos ou agentes computacionais, que utilizam-se da fase de recuperação, realizada individualmente ou em lotes, atendendo ao processamento humano e de máquina, sendo que esse último pode alimentar a coleta de outros CVDs, como o de repositórios digitais. No último caso, é necessário que o indexador de periódicos forneça um protocolo de comunicação ou API específica para a coleta de dados.

Neste trabalho são analisadas informações relativas às bases SciELO Brasil[3] e Scopus[4], disponíveis em suas páginas web.

A SciELO Brasil é uma biblioteca de periódicos científicos brasileiros, sediada na Universidade Federal de São Paulo (UNIFESP) e mantida pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Disponibiliza uma série de documentos textuais visando auxiliar os detentores interessados em submeter seu periódico à avaliação, embasada nos 'Critérios, política e procedimentos para a admissão e a permanência de periódicos científicos na Coleção SciELO Brasil' (SCIELO, 2017), que objetivam orientar o desenvolvimento da coleção tornando-se referencial para avaliação. Nesse sentido, destaca ainda o objetivo de contribuir para o desenvolvimento da pesquisa e para a visibilidade, credibilidade, uso e impacto, utilizando-se da indexação e da interoperabilidade nacional e internacional (SCIELO, 2018).

No que tange à coleta e recuperação, a SciELO Brasil armazena e disponibiliza, em acesso aberto, textos completos de periódicos científicos de qualquer área do conhecimento, desde que utilizem o processo de avaliação por pares e que apresentem desenvolvimento crescente quanto aos critérios de indexação.

[3] Disponível em: http://www.scielo.br/?lng=pt.
[4] Disponível em: https://www.scopus.com/home.uri.

Uma vez indexado, a etapa de coleta desse CVD é iniciada, cabe ao periódico o envio dos pacotes (fascículos e lotes de arquivos) utilizando File Transfer Protocol (FTP), em pasta designada para cada ação, nomeadamente entrega e correção; caso haja algum problema com o envio deve ser utilizado WeTransfer ou SendSpace. Os metadados dos artigos devem atender ao SciELO Publishing Schema (SPS), um conjunto de especificações para a marcação dos artigos utilizando eXtensible Markup Language (XML). Após o envio o tempo médio para disponibilização é de 10 a 15 dias corridos (SCIELO, 2018).

Apesar de se tratar de uma base indexadora, a SciELO Brasil disponibiliza uma página para o periódico, na qual, além dos metadados, fornece também os objetos digitais, atuando como um provedor de dados. Desta forma, há periódicos que possuem uma interface própria e sua página na SciELO Brasil, oferecendo dois pontos de acesso diferentes ao conteúdo. Assim, verificam-se algumas divergências entre as páginas dos periódicos e da SciELO Brasil, a exemplo da adoção de licenças de Direitos Autorais e do uso do Digital Object Identifier (DOI) da SciELO Brasil em detrimento da própria fonte de publicação.

A recuperação dos dados e objetos é possível pela interface SciELO Brasil, que oferece mecanismo de busca e elementos de navegação e, ainda por buscadores como Google e Google Acadêmico. Enquanto a recuperação por aplicações computacionais pode ser realizada pelo OAI-PMH.

Apesar da transitoriedade da indexação, a base não dispõe de informações acerca do descarte de dados dos periódicos que deixam de ser indexados pelo descumprimento aos critérios de avaliação.

Por outro lado, a Scopus, mantida pela Elsevier, “é o maior banco de dados de resumos e citações de publicações revisadas por pares do mundo, incluindo revistas científicas, livros e anais de conferências, cobrindo tópicos de pesquisa em todas as disciplinas técnicas e científicas [...]” (SCOPUS, 2018, tradução nossa). Todo o conteúdo indexado contém metadados e permite a integração com outras plataformas utilizando APIs disponíveis gratuitamente aos assinantes ou mediante aquisição de uma chave de acesso.

A Scopus atua como provedor de serviços, de forma que disponibiliza os metadados dos conteúdos indexados e *link* para acesso ao objeto digital na fonte original de publicação. Apesar disso, não há informações claras acerca do armazenamento, tampouco do descarte dos dados.

Os dados armazenados podem ser recuperados via mecanismo de busca e elementos de navegação disponíveis na interface. São fornecidos ainda alguns recursos como alertas de pesquisa, busca avançada por fonte de publicação, listas personalizadas pelo usuário para uma seção específica de pesquisa, verificação de artigos que citam o material selecionado e visualização da lista de referências incluídas no artigo.

No que tange às APIs, há outros diferenciais, como a classificação de assuntos associados ao conteúdo, *links* para recursos relacionados (perfil de autor, afiliação), recuperação por afiliação institucional do(s) autor(es) e recuperação pelo perfil do autor. Para instituições que adquirem a chave de acesso, além dos recursos anteriores, há suporte para padrões de metadados e especificações (W3C) que favorecem a integração com aplicativos ou *web services*.

2.3 Ciclo de Vida de Dados em Repositórios Digitais

O movimento de acesso aberto trouxe como uma das prerrogativas a disponibilidade livre e irrestrita de resultados de pesquisa. Nesse sentido, aconteceram iniciativas de disponibilização de *preprints,* como forma de apresentar as contribuições aos pares, a exemplo do ArXiV (MUELLER, 2006). Por outro lado, houve o advento de repositórios digitais, em sua maioria vinculados a instituições de pesquisa, com objetivos como a disponibilização e ampliação da visibilidade de resultados de pesquisa publicados em outras fontes, bem como a preservação para acesso a longo prazo.

Para esta pesquisa, fez-se a análise do DSpace, software mais utilizado para a disponibilização de repositórios digitais no Brasil. O DSpace é uma ferramenta desenvolvida inicialmente pela MIT Libraries e Hewlett-Packard (HP), atualmente mantida e distribuída pelo DuraSpace, uma organização independente e sem fins lucrativos. Desenvolvido em linguagem JAVA, utilizando banco de dados SQL, o software é distribuído utilizando uma licença Berkeley Software Distribution (BSD) e sua documentação uma licença Creative Commons, constituindo-se em uma iniciativa de código aberto para que seja melhorado pela comunidade, cujas contribuições podem ser compartilhadas (DURASPACE, 2018).

Ao considerarmos o CVD de repositórios digitais gerenciados pelo DSpace, o fluxo é iniciado pela coleta. Nesse aspecto, na Budapest Open Access Iniciative (BOAI) (2002) definiu como uma das estratégias de acesso aberto o autoarquivamento em repositórios digitais, pelo

pesquisador, das suas publicações. Transcorridos 10 anos dessa experiência, foram adicionadas novas recomendações (BUDAPEST OPEN ACCESS INITIATIVE, 2012), dentre as quais destaca-se a adoção de licenças de acesso aberto para que os conteúdos publicados em periódicos possam ser reutilizados e o incentivo aos repositórios para a coleta de artigos e depósito em outras fontes, o que pode ser utilizado para o povoamento de repositórios digitais.

Tal perspectiva foi adotada pelo processo de coleta automática de registros disponíveis nos ambientes digitais das fontes publicadoras (VIDOTTI *et al.* 2015; VIDOTTI *et al.* 2016), visando otimizar o povoamento dos repositórios, desonerando o pesquisador. Qualquer desses processos é que alimenta a etapa de coleta no ciclo de vida dos repositórios digitais, conforme explicitado na Figura 1.

Para a coleta de dados, o DSpace, quando da instalação, apresenta uma estrutura de metadados em Dublin Core, sendo possível customizá-lo para estabelecer atributos por tipo de documento, e definir quais desses são obrigatórios e repetitivos; além disso, após a customização, é necessário registrar os metadados, via interface, com o perfil de administrador do sistema

Vale destacar que o uso de um esquema de metadados internacionalmente aceito é imprescindível para que os dados sejam interoperáveis. Nesse sentido, quando a coleta é alimentada por humanos, a representação do objeto digital a ser armazenado é realizada pelo(s) autor(es). Por outro lado, no processo de coleta automática, caso o repositório digital utilize um perfil de aplicação contendo um padrão de metadados ou atributos divergentes da fonte original de publicação, pode ser necessário compatibilizar automaticamente campos e, ainda, realizar tratamento manual dos dados antes da entrada no base de dados.

No caso de depósito automático, o DSpace oferece a possibilidade de entrada de dados utilizando o menu da própria interface do sistema no perfil de Administrador, ou ainda a utilização de linha de comando, *web service* ou Representational State Transfer (REST)[5].

Da mesma forma que indicado na seção relativa aos periódicos, deve-se atender às boas práticas de publicação de dados na web preconizadas pelo World Wide Web Consortium (W3C) (LÓSCIO; BURLE; CALEGARI, 2017) para atribuir contexto semântico ao objeto

[5] O DSpace disponibiliza informações acerca da importação de dados em sua Wiki.

disponibilizado, favorecendo o processo de armazenamento, interpretação, recuperação, encontrabilidade e ligação dos dados.

O DSpace não limita o formato do objeto digital submetido, contudo é possível configurar os formatos e tamanhos de arquivos aceitos. Uma característica do sistema é a disponibilização web de objetos para ampliação de uso, de forma que não oferece a opção de versionamento de arquivos, estando o *bitstream* depositado disponível na interface de recuperação. Nesse sentido, o sistema permite a entrada de objetos e fornece a opção de acesso em níveis, tornando passível de configuração de acesso todos os arquivos depositados, quer sejam para acesso público imediato, acesso restrito a uma comunidade de usuários ou por tempo determinado. Tal característica é relevante por solucionar questões relacionadas aos direitos autorais, aos processos de registro de patentes ou outra condição legal imposta ao objeto; de igual maneira, é possível determinar se os itens que apresentam restrições devem ter os metadados exibidos publicamente ou a um grupo de usuários.

A base de dados do DSpace pode ser MySQL ou PostgreSQL, cabendo à instituição mantenedora (detentor) a definição, considerando a infraestrutura disponível. Os dados armazenados podem ser exibidos a partir da configuração do protocolo Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), que possibilita a recuperação da informação em pacotes por aplicações computacionais, em diferentes padrões de metadados, como Dublin Core (DC), Metadata Encoding and Transmission Standard (METS), Metadata Object Description Schema (MODS), Machine-Readable Cataloging (MARC). Permite ainda a coleta por motores de busca, a exemplo do Google e Google Acadêmico. De igual maneira, o sistema disponibiliza interface que possibilita a recuperação da informação por humanos, por meio da navegação e/ou do mecanismo de busca.

No CVD de repositórios digitais, o detentor é a instituição mantenedora, que determina o processo de negócios, bem como as formas de armazenamento, descarte e recuperação, que poderão ser por ele operacionalizadas ou delegadas a um intermediário, que pode ser parte da infraestrutura ou uma estrutura parceira. No caso do DSpace, em primeira instância, o intermediário é o DuraSpace, enquanto desenvolvedor do sistema, podendo esse papel ser compartilhado com outros atores que tenham realizado alterações no código fonte do software. Os referenciados são o(s) pesquisador(es) que assumem o papel de autor(es) do(s) trabalho(s) publicado(s), a(s) instituição(ões) de afiliação, a(s) agência(s) de fomento que apoiou(aram) o desenvolvimento da pesquisa e, os dados do periódico no qual

o artigo foi publicado. Os usuários podem ser humanos ou agentes computacionais, que utilizam-se da fase de recuperação, realizada individualmente ou em lotes, atendendo a processamento humano e de máquina, sendo que esse último pode alimentar a coleta de outros CVDs.

No que tange à segurança, é possível considerar como sigilosos os próprios objetos, que podem ter acesso restrito, conforme descrito anteriormente.

3 DISCUSSÕES

O campo informacional estudado apresenta três CVDs, periódicos, indexadores de periódicos e repositórios digitais, além de dois atores, o pesquisador e a comunidade científica (Figura 1). A análise realizada destaca que, os três CVDs se relacionam, bem como possuem os mesmos atores envolvidos, contudo, os dados e objetos que os circundam estão dependentes do processo de coleta do CVD de periódicos, cuja alimentação é realizada pelo pesquisador, sendo esse considerado o principal ator.

A estruturação adequada do CVD de periódicos é imprescindível, uma vez que é a partir da sua recuperação que os demais ciclos alimentarão suas etapas de coleta e, de igual maneira, haverá recuperação por parte dos usuários. Com isso, os fatores: privacidade, integração, qualidade, direitos autorais, disseminação e preservação, apresentados por Sant'Ana (2013; 2016), também são de relevante tratamento nesse CVD.

Nesse prisma compete ao detentor e ao intermediário algumas definições que afetam diretamente os fatores elencados anteriormente, para os quais, o pesquisador, enquanto principal ator no campo informacional deve ter clareza das decisões a serem tomadas ao alimentar a coleta do ciclo, considerando que, após a entrada dos dados no sistema o pesquisador deixa de ter domínio sobre eles.

Nesta perspectiva, cumpre-se ressaltar que a manipulação e a exibição dos dados alimentados na coleta passam a ser prerrogativa do detentor e, a possibilidade de coleta e exibição desses dados por motores de busca ampliam a coleta por múltiplas fontes, dificultando o rastreamento.

Desta forma são abordados os fatores que geram impacto ao pesquisador, nomeadamente privacidade e direitos autorais e, dada a similaridade dos CVDs apresentados nesse estudo, utiliza-se como base o ciclo de periódicos para generalizar a discussão, trazendo para a análise questões concernentes aos demais ciclos quando pertinentes.

É notório que o advento das tecnologias privilegia a circulação indiscriminada de dados em sistemas proprietários ou de código aberto, administrados por organizações com diferentes finalidades, em redes privadas ou na web.

No Brasil, tal situação foi abordada na legislação apenas em 2014, quando da promulgação da Lei nº 12.965 (BRASIL, 2014) que "estabelece princípios, garantias, direitos e deveres para o uso da internet no Brasil". Tal ordenamento jurídico, em seu artigo terceiro aborda a proteção da privacidade, considerando sobretudo os dados pessoais, que passou a ser regulado no ano de 2018 pela Lei nº 13.709 (BRASIL, 2018), alterada antes de entrar em vigor, em 2019, pela Lei nº 13.853 (BRASIL, 2019), passando a vigorar 24 meses após a sua publicação original, ou seja, em agosto de 2020.

O referido ordenamento jurídico, cuja ementa designa como a Lei Geral de Proteção de Dados Pessoais (LGDP) "dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural" (BRASIL, 2018).

Cumpre destacar que a legislação (BRASIL, 2018) prevê atores envolvidos no ciclo de vida de dados, embora utilize-se de terminologia divergente de Sant'Ana (2016) e, atribui ao que designa titular (referenciado, para Sant'Ana) a titularidade dos seus dados pessoais, cabendo a ele as prerrogativas de uso e tratamento, estabelecendo, inclusive, suas formas. Conquanto também estabelece papéis para outros atores, a exemplo de detentores e intermediários (na lei considerados controlador e operador, respectivamente) no que tange à coleta, ao armazenamento, ao tratamento e à segurança dos dados.

A partir da análise da legislação apresentada, é possível afirmar que, embora seja prerrogativa do referenciado a permissão de coleta e de exclusão dos seus dados pessoais da base de dados, quem de fato manipula e define as formas de tratamento dos dados é o detentor.

Sant'Ana (2016) afirma que qualquer referenciado em uma base de dados deveria ter o direito a solicitar a retirada e evoca o direito ao esquecimento. Contudo, clarifica-se, a partir da análise, a ausência de controle desse ator em relação ao CVD após a fase de coleta.

Destaca-se ainda que a legislação brasileira aborda apenas a questão dos dados pessoais e que não há menção a outros dados.

A esse respeito, o CVD de periódicos, utilizando o OJS requer o cadastro de dados pessoais dos autores, muito embora não faça a requisição de documentos pessoais ou dados sensíveis. O CVD de repositórios digitais requer cadastro apenas quando a coleta é realizada por autoarquivamento. No campo informacional descrito, os dados pessoais dos autores não são intercambiáveis, uma vez que os demais CVDs coletam apenas os dados da publicação.

No que tange aos direitos autorais, no Brasil, a Lei nº 9.610 (BRASIL, 1998) estabelece a natureza dúplice dos direitos, que nascem com o autor, que detém os direitos morais e a titularidade dos direitos patrimoniais sobre a obra. A primeira forma de direito protege a pessoa, a quem deve ser reconhecida a autoria perpetuamente, sem a possibilidade de qualquer tipo de transferência, cessão ou concessão; enquanto a segunda relaciona-se à materialização da ideia expressa pelo autor, ou seja a obra. A titularidade dos direitos sobre a obra é, em primeira instância do autor, contudo, pode ser transferida a pessoas físicas ou jurídicas mediante especificação contratual.

Nesse sentido, ainda é comum que os periódicos científicos solicitem aos autores a cessão dos direitos autorais, especificação disponível nas diretrizes para autores e como condição de submissão do manuscrito durante a fase de coleta. Vale destacar que, caso o autor concorde com a cessão da titularidade dos direitos autorais, ele deixa de ter a prerrogativa de tomar decisões acerca da publicação, ficando essa sob responsabilidade do titular dos direitos, nesse caso, o periódico. Cabe assim, em primeira análise, ao autor a decisão de publicar em periódicos que condicionem a coleta à transferência de titularidade.

Uma vez publicado um resultado de pesquisa, é possível determinar, via metadados no OJS se o artigo possui todos os direitos reservados por *copyright* ou ainda alguns direitos reservados, cuja especificação apresenta-se pelo uso de uma licença aberta, sendo a mais usual a Creative Commons[6]. Esse dado é de fundamental importância para a fase de recuperação, pois determinará quais usos do objeto podem ser realizados, seja por humanos ou aplicações computacionais, sem requerer a autorização do titular dos direitos autorais.

A crescente disponibilização de periódicos em ambientes digitais, com acesso gratuito construiu uma percepção errônea do acesso aberto, fato esse embasado na matéria jurídica explicitada. Assim, no momento da recuperação, seja por humanos ou agentes computacionais, é imprescindível certificar-se das condições de acesso e uso do objeto digital.

[6] Para mais informações sobre as licenças, consulte a página. Disponível em: https://br.creativecommons.org/licencas/. Acesso em: 28 nov. 2018.

No que tange aos direitos autorais, é comum encontrarmos divergências entre as informações disponíveis em diferentes áreas no website do periódico e ainda entre periódico e indexador. Infere-se que a discrepância de informações no periódico sinaliza desconhecimento da legislação autoral bem como das formas de licenciamento, o que acarreta prejuízos à publicação, colocando em risco a instituição mantenedora e o autor. Ao considerarmos que a entrada dos dados é realizada no CVD de periódicos, é esse quem deve definir as condições de publicação, de acesso e uso aos objetos publicados. Conquanto, há indexadores de periódicos que definem uma licença Creative Commons mandatória, que condiciona a entrada do registro na base de dados. Contudo, não há qualquer prerrogativa por parte dos detentores de CVDs correlacionados, cabendo as definições apenas e tão somente ao titular dos direitos que, nesse caso, pode ser o pesquisador ou o detentor do CVD de periódicos.

O OJS possibilita a configuração da política de acesso aos conteúdos publicados, mantendo o *copyright* ou pela atribuição de uma licença Creative Commons. Independente da forma de acesso determinada pelo periódico é altamente recomendável que haja menção à condição de acesso e uso nos metadados do registro, de forma a exibi-lo para humanos e agentes computacionais, favorecendo o usuário, para que não infrinja à legislação vigente, e, os demais CVDs do campo informacional que podem utilizar-se desse dado nas coletas. Recomenda-se ainda o registro da licença em um diretório de políticas de direitos autorais, a exemplo do SHERPARomeo[7] para publicações internacionais e do Diadorim[8] para as nacionais.

Vale destacar que a fonte de publicação é o periódico a quem compete o estabelecimento dos requisitos iniciais, a exemplo dos metadados utilizados para a representação do objeto digital publicado, bem como das definições concernentes aos direitos autorais. Isso posto, o indexador de periódicos deve apenas atuar como ferramenta de ampliação de visibilidade e de chancela à publicação, não cabendo a ele definições ou alterações dos dados oriundos do periódico. Nesse sentido, cumpre destacar que a SciELO Brasil adota uma licença de direitos autorais do tipo Creative Commons específica para todo o conteúdo indexado e, considerando a coexistência de uma página própria da revista, podem haver divergências nas licenças adotadas. Assim, é imprescindível a interpretação da

[7] Disponível em: http://www.sherpa.ac.uk/romeo/index.php?la=pt.
[8] Disponível em: http://diadorim.ibict.br/.

legislação de direitos autorais, de forma ao entendimento de quem compete o licenciamento do material.

De igual maneira, os repositórios digitais devem utilizar a mesma forma de direito quando da coleta dos dados, sob pena de colocar em risco a sua reputação. "A confiabilidade dos dados é condição *sine qua non* para que um dado seja útil." (SANT'ANA, 2016, p. 125). Reforça-se assim a relevância de manter dados de proveniência, bem como da condição de acesso e uso de forma idêntica à fonte no repositório digital.

4 CONSIDERAÇÕES FINAIS

O presente trabalho apresenta os ciclos de vida de dados e os atores envolvidos no processo de publicação e acesso à produção científica; as formas de coleta, armazenamento e recuperação de dados que compõem o campo informacional descrito de modo a fornecer subsídios para o reuso de dados dos ciclos apresentados em diferentes contextos; identifica como principal ator o pesquisador e descreve os fatores que o impactam e são impactados por ele.

Dessa forma, discute privacidade e direitos autorais, considerando a legislação brasileira vigente e reforça o protagonismo do pesquisador que, em primeira instância é quem insere seus dados e objetos digitais no CVD de periódicos e de repositórios digitais (pelo autoarquivamento), e a quem compete definir a melhor forma de tratamento dos dados, quer seja na privacidade requerida ou na definição da manutenção ou transferência da titularidade dos direitos autorais patrimoniais do conteúdo para a fonte publicadora, bem como das permissões de acesso e uso.

Nesse momento, o pesquisador possui todas as prerrogativas da decisão, que devem ser tomada de forma consciente e, em atendimento aos seus anseios. Ao passo que reforça a impotência do indivíduo após a disponibilização dos seus dados na fase de coleta do CVD, cujo armazenamento e descarte passam a ser gerenciados pelo detentor, sobre o qual o pesquisador, enquanto referenciado, não exerce nenhum tipo de influência. Além disso, vale destacar a possibilidade de que os dados armazenados sejam coletados por outros CVDs, sobre os quais sequer o detentor do CVD no qual o referenciado os inseriu tem domínio, dificultando a rastreabilidade.

REFERÊNCIAS

BRASIL. Lei nº 12.965, de 23 de abril de 2014. **Diário Oficial da União**, Brasília, DF, 24 abr. 2014.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. **Diário Oficial da União**, Brasília, DF, 15 ago. 2018.

BRASIL. Lei nº 13.853, de 8 de julho de 2019. **Diário Oficial da União**, Brasília, DF, 9 set. 2019.

BRASIL. Lei nº 9.610, de 19 de fevereiro de 1998. **Diário Oficial da União**, Brasília, DF, 20 fev. 1998. Seção 1. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L9610.htm. Acesso em: 30 nov. 2018.

BUDAPEST OPEN ACCESS INITIATIVE. **Dez anos da iniciativa de Budapeste em acesso aberto**: a abertura como caminho a seguir. 2012. Disponível em: https://www.budapestopenaccessinitiative.org/boai-10-translations/portuguese-brazilian-translation. Acesso em: 01 nov. 2018.

BUDAPEST OPEN ACCESS INITIATIVE. **Iniciativa de Budapeste pelo acesso aberto**. 2002. Disponível em: https://www.budapestopenaccessinitiative.org/translations/portuguese-translation. Acesso em: 01 nov. 2018.

CAMARGO, L. S. de A. de; VIDOTTI, S. A. B. G. **Arquitetura da informação**: uma abordagem prática para o tratamento de conteúdo e interface em ambientes informacionais digitais. Rio de Janeiro: LTC, 2011.

CÔRTES, P. L. Considerações sobre a evolução da ciência e da comunicação científica. *In*: POBLACION, D. A.; WITTER, G. P.; SILVA, J. F. M. da (org). **Comunicação & produção científica**: contexto, indicadores e avaliação. São Paulo: Angelara, 2006.

DURASPACE. **DSpace**. Disponível em: https://duraspace.org/dspace/. Acesso em: 01 nov. 2018.

GARVEY, W. D. **Communication**: the essence of science. Oxford: Pergamon Press, 1979.
KLING, R.; CALLAHAN, E. Electronic Journals, the Internet, and Scholarly Communication. *In*: CRONIN, Blaise (ed.). **Annual Review of Information Science and Technology**: Information Today, v. 37, p. 127-77, 2003.

LEITE, F. C. L. **Como gerenciar e ampliar a visibilidade da produção científica brasileira**: repositórios institucionais de acesso aberto. Brasília: Ibict, 2009. Disponível em: http://livroaberto.ibict.br/bitstream/1/775/4/Como%20gerenciar%20e%20ampliar%20a%20visibilidade%20da%20informa%C3%A7%C3%A3o%20cient%C3%ADfica%20brasileira.pdf. Acesso em: 30 nov. 2018.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. (ed.). **Data on the web best practices**. 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 26 abr. 2018.

LYNCH, C. A. Institutional repositories: essential infrastructure for scholarship in the digital age. **ARL Bimonthly Report**, Washington, n. 226, 2003. Disponível em: http://old.arl.org/resources/pubs/br/br226/br226ir~print.shtml. Acesso em: 17 dez. 2008.

MEADOWS, A. J. **A comunicação científica**. Brasília: Briquet de Lemos, 1999.

MUELLER, S. P. M. A comunicação científica e o movimento de acesso livre ao conhecimento. **Ciência da Informação**, [s.l.], v. 35, n. 2, p. 27-38, maio/ago. 2006.

PUBLIC KNOWLEDGE PROJECT. **Open Journal System**. Disponível em: https://pkp.sfu.ca/ojs/. Acesso em: 01 nov. 2018a.

PUBLIC KNOWLEDGE PROJECT. **User guides, developer documentation and publishing tips for all of the Public Knowledge Project's software**. Disponível em: https://docs.pkp.sfu.ca/#appojs3. Acesso em: 01 nov. 2018b.

SANT'ANA, R. C. G. Ciclo de vida dos dados e o papel da ciência da informação. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Florianópolis. **Anais eletrônicos** [...]. Florianópolis, 2013.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da Informação. **Informação & 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. Acesso em: 14 set. 2018.

SCIELO. **Critérios, política e procedimentos para a admissão e a permanência de periódicos científicos na Coleção SciELO Brasil**. 2017. Disponível em: http://www.scielo.br/avaliacao/Criterios_SciELO_Brasil_versao_revisada_atualizada_outubro_20171206.pdf. Acesso em: 10 dez. 2018.

SCOPUS. **Sobre.** 2018. Disponível em: https://www.scopus.com/search/form.uri?display=basic. Acesso em: 10 dez. 2018.

VIDOTTI, S. A. B. G. *et al*. Coleta automática para o povoamento de repositórios digitais: conversão de registros utilizando XSLT. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. **Anais eletrônicos** [...]. Salvador: UFBA, 2016.

VIDOTTI, S. A. B. G. *et al*. Reutilização de metadados para o povoamento de um repositório institucional: procedimentos aplicados no Repositório Institucional UNESP. *In*: INTERNATIONAL CONFERENCE ON Dublin Core & METADATA APPLICATIONS (DC-2015), 15., 2015, São Paulo. **Proceedings** […]. São Paulo: Unesp, 2015.

WEITZEL, S. da R. Fluxo da informação cientifica. *In*: POBLACION, D. A.; WITTER, G. P.; SILVA, J. F. M. da (orgs.). **Comunicação & produção científica**: contexto, indicadores e avaliação. São Paulo: Angelara, 2006.