Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação
ISSN: 1518-2924
adilson.pinto@ufsc.br
Universidade Federal de Santa Catarina
Brasil

RAUTENBERG, Sandro; Cassiana BURDA, Alessandra; de SOUZA, Lucélia
Um workflow para compartilhamento de dados científicos primários baseado em dados abertos conectados
Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, vol. 23, núm. 53, 2018, Setembro-, pp. 110-123
Universidade Federal de Santa Catarina
Brasil

DOI: https://doi.org/10.5007/1518-2924.2018v23n53p110

Disponível em: https://www.redalyc.org/articulo.oa?id=14762417011

ARTIGO

Recebido em:
28/11/2017

Aceito em:
18/04/2018

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 23, n. 53, p.110-123, set./dez., 2018. ISSN 1518-2924. DOI: 10.5007/1518-2924.2018v23n53p110

Um workflow para compartilhamento de dados científicos primários como Dados Abertos Conectados

A workflow for sharing primary data as Linked Open Data

Sandro RAUTENBERG (srautenberg@unicentro.br) *
Alessandra Cassiana BURDA (alessandra.burda@gmail.com) **
Lucélia de SOUZA (lucelia@unicentro.br)***
* Doutor em Engenharia e Gestão do Conhecimento pela Universidade Federal de Santa Catarina.
** Bacharel em Ciência da Computação pela Universidade Estadual do Centro-Oeste.
*** Doutora em Ciência da Computação pela Universidade Federal do Paraná.

Resumo
Investiga a automação do processo de publicação de dados abertos científicos na Web de Dados. Metodologicamente, o trabalho é baseado no ciclo de vida Linked Data Lifecycle e suas tecnologias. Como resultado, apresenta-se um *Workflow* para compartilhar conjuntos de dados primários no contexto da Cientometria, preservando os históricos dos índices Qualis, SJR e SNIP na Web de Dados (período 2005-2016). Conclui-se que o *Workflow* é um instrumento tecnológico importante na preservação de dados científicos primários, suportando as pesquisas científicas quanto ao reuso de recursos e à reprodutibilidade de resultados.
Palavras-chave: Dados Abertos Conectados. *Workflow*. *Workflow* para Dados Abertos Conectados. Dados Primários.

Abstract
We investigate the automation of the process for publishing scientific open data on the Web of Data. This work is based on the Linked Data Lifecycle and its technologies. As a result, a Workflow is established for sharing primary datasets in the Scientometric domain, preserving historical records of the Qualis, SJR e SNIP indexes on the Web of Data (2005-2016). As conclusion, we stand that this establishment is an important technological instrument for digital preservation of scientific data and can support scientific researches, considering the reuse of resources and the result's reproducibility.
Keywords: Linked Open Data. Workflow. Workflow for Linked Open Data. Raw Data.

v. 23, n. 53, 2018.
p. 110-123
ISSN 1518-2924

1 INTRODUÇÃO

A Web de Dados é uma estrutura de vanguarda para o compartilhamento de Dados Abertos Conectados em escala global, propiciando a exploração de recursos de dados nos mais variados domínios. Nesse ambiente informacional, o tratamento dos recursos disponibilizados enseja um campo profícuo de atuação da Ciência da Informação. Em poucas palavras, na Web de Dados, a manutenção de recursos de dados envolve processos amplamente discutidos na Ciência da Informação, no tocante ao planejamento das atividades de gestão do ciclo de vida dos Dados Abertos Conectados.

Além de um campo de atuação da Ciência da Informação, a Web de Dados também pode ser uma plataforma de apoio às disciplinas desta área. Os estudos deste domínio podem se beneficiar da Web de Dados, por acessar os recursos de dados aderentes ao objeto de pesquisa. Ou seja, da Web de Dados, se obtém os Dados Abertos Conectados, os quais são convertidos em dados para subsidiar as pesquisas. Desta forma, minimizam-se os esforços nos processos de aquisição, triagem, tratamento e utilização de dados primários. Por isso, é pertinente estabelecer medidas quanto à manutenção de Dados Abertos Conectados, fomentando a base informacional da Web de Dados.

Baseando-se nas assertivas anteriores, este trabalho aborda os processos de compartilhamento de dados primários como Dados Abertos Conectados. Pontualmente, investiga-se um *Workflow* para auxiliar a manutenção de dados primários cientométricos na Web de Dados. Configurando-se como uma pesquisa aplicada, o *Workflow* é empregado e verificado na preservação dos históricos dos índices Qualis, SJR (SCImago Journal & Country Rank) e SNIP (Source Normalized Impact per Paper).

Para discutir o *Workflow* proposto, este artigo ainda compreende as seguintes seções: **(i)** a fundamentação teórica, a qual discorre sobre os conceitos Dados Abertos Conectados e *Workflow*; **(ii)** o procedimento metodológico, que aponta o ciclo de vida para Dados Abertos Conectados como inspiração para a definição dos passos do *Workflow* desenvolvido para preservação de Dados Abertos Conectados, as ferramentas utilizadas na execução computacional do *Workflow* proposto e os conjuntos de dados abertos primários utilizados; **(iii)** a definição do *Workflow* proposto e seus passos constituintes; **(iv)** a verificação do *Workflow*, apresentando sua execução na preservação dos índices cientométricos; e por fim **(v)** as considerações finais, abordando os resultados alcançados e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A base constitutiva deste trabalho é alinhada ao entendimento dos conceitos Dados Abertos Conectados e *Workflow*. A seguir, os referidos conceitos são discutidos de forma interdisciplinar.

2.1 Dados Abertos Conectados

Em suma, os dados são classificados como Dados Abertos Conectados quando disponibilizados na web para seu livre reuso (OPEN KNOWLEDGE INTERNATIONAL, 2017; HEATH; BIZER, 2011). Neste sentido, esses recursos são publicados de acordo com licenças abertas, possibilitando que sejam reutilizados sem restrições, por pessoas ou agentes de software para explorar sua capacidade informacional em diversos domínios.

Constitutivamente, a percepção de Dados Abertos Conectados é vinculada a dois entendimentos: (a) o que são dados abertos; e (b) como os dados são conectados. Os dados são considerados abertos quando "podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras" (OPEN KNOWLEDGE INTERNATIONAL, 2017). Ressalta-se que os dados abertos são classificados de acordo com seu nível de abertura e sua conexão a outros dados. Denominada 5-Estrelas, a referida classificação é organizada conforme segue (5-STAR, 2017):

> **1ª Estrela** - é atribuída aos conjuntos de dados publicados sob uma licença aberta (*Open License* - OL), entretanto, em um formato proprietário. Os dados somente podem ser manipulados (lidos, visualizados ou impressos) por determinados softwares. Como exemplo, tem-se os dados publicados no formato *Portable Document Format* (PDF).

2ª Estrela - é conferida à publicação de conjuntos de dados estruturados legíveis por máquinas (*Readable Machine* - RE). A esse nível, os dados são processados por softwares proprietários e podem ser exportados em outros formatos, facilitando os procedimentos de manipulação. As planilhas eletrônicas no formato *eXceL Spreadsheet* (XLS - formato de planilha eletrônica da Microsoft) são classificadas neste nível.

3ª Estrela - é concedida aos conjuntos de dados publicados em algum formato aberto (*Open Format* – OF), como por exemplo, *Comma Separated Value* (CSV) ou *JavaScript Object Notation* (JSON). Salienta-se que a manipulação dos dados no formato aberto não necessita o uso de um software proprietário.

4ª Estrela - é designada à utilização dos Identificadores Uniforme de Recursos (*Uniform Resource Identifier* - URI) para rotular os dados, permitindo que os usuários criem ligações e façam o reuso dos dados disponibilizados.

5ª Estrela - é atribuída aos conjuntos de dados que são conectados (*Linked Data* - LD) a outros dados em uma infraestrutura de rede. Isso permite a navegação entre dados e a descoberta de informação. Dessa forma, acrescenta-se valor aos recursos de dados ao promover uma contextualização mais ampliada.

Considerando a classificação anterior, a união de dados abertos e dados conectados é estabelecida ao se atingir a 5ª Estrela. Isso representa o ideal de publicação de dados na Web de Dados. Ou seja, na Web de Dados, os dados abertos estão conectados a outros dados distribuídos na rede, constituindo os Dados Abertos Conectados. Ressalta-se que essa união constitui a base informacional de um imenso grafo de recursos de dados, a Web de Dados. Neste sentido, a publicação de Dados Abertos Conectados tem como objetivo usar a arquitetura da web para compartilhar dados estruturados em uma escala global. Assim, incentiva-se o reuso do conjunto de dados universal por diferentes pessoas e aplicações.

Para fomentar a Web de Dados, ao publicar novos recursos em sua plataforma global, este trabalho incrementa os níveis de abertura de alguns conjuntos de dados primários (os índices cientométricos Qualis, SJR e SNIP). Considerando que originalmente os referidos índices cientométricos se apresentam em formatos proprietários nas 1ª e 2ª Estrelas (PDF ou XLS), estes tiveram seu nível de abertura elevado ao patamar da 5ª Estrela. Desta maneira, facilita-se o reuso desses recursos em demais pesquisas cientométricas na forma de Dados Abertos Conectados. Tal empreendimento é promovido mediante a definição e utilização de um *Workflow* conforme os preceitos discutidos a seguir.

2.2 *Workflow* para Dados Abertos Conectados

A criação, a manutenção e o compartilhamento de Dados Abertos Conectados são atividades complexas que requerem o uso substancial de recursos e de tecnologias para preservar os conjuntos de dados na Web de Dados. Os esforços relacionados a essas atividades devem ser planejados, possibilitando sua execução sistemática em conformidade às melhores práticas de publicação de Dados Abertos Conectados. Neste enredo, insere-se o conceito de *Workflow*.

Workflow, ou fluxo de trabalho, é um conceito interdisciplinar, estudado/empregado em vários domínios. Na Ciência da Informação, segundo o Dicionário de Biblioteconomia e Arquivologia (CUNHA; CAVALCANTI, 2008, p. 170), o conceito *Workflow* envolve a sinergia de diversas partes (usuários, organização, sistemas, documentos, processos, ações, controle e fluxo de dados) conforme entendido a partir de sua definição:

> Sistema de encaminhamento automático de documentos para os usuários ligados a uma determinada organização [...] O sistema geralmente inclui as ações a serem realizadas, indicações sobre o controle e o fluxo de dados, as pessoas autorizadas a executá-las e a descrição do ambiente organizacional. Na prática, o conceito refere-se ao fluxo do processo do negócio dentro de uma determinada organização [...].

Num âmbito mais técnico e em consonância à definição anterior, The Workflow Management Coalition (HOLLINGSWORTH, 2017, tradução nossa) conceitua *Workflow* como:

> [...] a automação de procedimentos em que documentos, informações ou tarefas são repassadas entre agentes de acordo com um conjunto definido de regras para alcançar ou contribuir com um objetivo geral.

Metodologicamente, um *Workflow* é constituído de um conjunto de passos, que organizados cronologicamente e, geralmente, usa Tecnologias de Informação e Comunicação para transformar insumos(s) durante sua execução. Em suma, o estabelecimento de um *Workflow* visa a sistematização de um processo (de negócio, de gestão de dados, por exemplo), tornando-o mais eficiente na produção de um resultado desejado (FERNANDES, 2012).

No âmbito deste trabalho (representado na Figura 1), para preservar dados na Web de Dados, um *Workflow* é constituído para orquestrar a execução de uma sequência de passos no tratamento desses recursos. A cada passo, um conjunto de dados de entrada é transformado em um conjunto de dados de saída, com o uso de uma ferramenta computacional de acordo com os parâmetros de configuração. Ao final, um conjunto de dados aderente aos preceitos de Dados Abertos Conectados deve ser produzido.

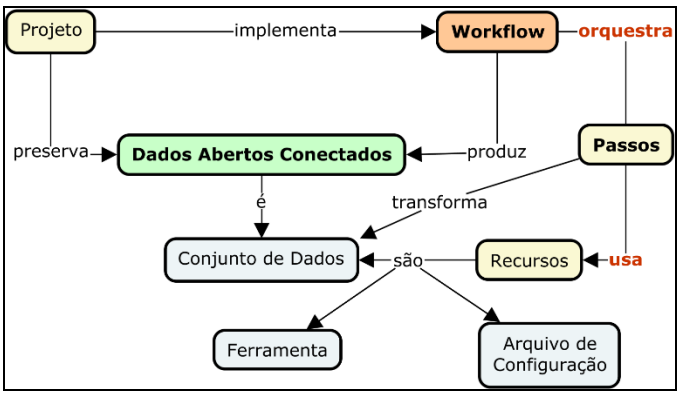

Figura 1: Mapa conceitual do *Workflow* para Dados Abertos Conectados.
Fonte: Dados da Pesquisa, 2017.

Diante as assertivas anteriores, o desenvolvimento de um *Workflow* para Dados Abertos Conectados deve privilegiar:

- **Requisito 1 - Planejamento** – a capacidade de descrever a estratégia de transformação de dados primários em Dados Abertos Conectados. Tal requisito é alcançado mediante a especificação e a ordenação temporal de um conjunto de passos. Tecnologicamente, em cada passo deve-se definir: (i) qual ferramenta computacional utilizar; (ii) a configuração da execução da referida ferramenta; (iii) o conjunto de dados de entrada; e (iv) o conjunto de dados de saída.
- **Requisito 2 – Execução –** é a capacidade de executar o *Workflow* planejado. Isso envolve um ambiente controlado, no qual as ferramentas computacionais, os arquivos de configuração e os conjuntos de dados devem ser disponibilizados. As ferramentas devem ser cronologicamente orquestradas de modo a transformar os conjuntos de dados de entrada em conjuntos de dados de saída a cada passo do *Workflow*. Ao final da execução, é desejável que, a partir de um conjunto original de dados primários, obtenha-se uma versão dos dados codificada em conformidade ao quinto nível da classificação 5-Estrelas.
- **Requisito 3 – Evolução dos dados –** dado um projeto de preservação de Dados Abertos Conectados, é a capacidade de atualizar o conjunto de dados ao longo do tempo, à medida que novos dados são gerados/refinados.

Ao atender os requisitos apresentados anteriormente, para publicar dados científicos primários na Web de Dados, um *Workflow* para Dados Abertos Conectados contribui na:

- **Reusabilidade de Dados**. Aumentam-se as chances de reutilização dos dados compartilhados na Web de Dados (W3C, 2017) em diferentes pesquisas, minimizando os esforços nas tarefas de coleta de dados.

• **Preservação Digital da Informação Científica**. Na perspectiva apontada por Arellano (2008), facilita o gerenciamento da informação científica ao longo do tempo, por meio da aplicação de estratégias de armazenamento, manutenção e acesso dos recursos digitais.
• **Reprodutibilidade de Resultados**. Permite a recuperação, a manipulação automatizada de dados, a integração dos dados entre as fontes distintas e a interoperabilidade de dados mediante a compatibilização de formatos (RDF, JSON, CSV, XLS, entre outros), contribuindo nos procedimentos metodológicos de pesquisas científicas.

Tais contribuições são mais bem percebidas na seção "5 Verificação do *Workflow*", onde são apresentados os passos realizados para a publicação dos índices cientométricos Qualis, SJR e SNIP na Web de Dados. Para tanto, a seguir são discutidas as bases metodológicas que sustentam esse empreendimento.

3 PROCEDIMENTO METODOLÓGICO

Nesta seção são apresentados: o ciclo de vida denominado Linked Data Lifecycle (AUER, 2014); as ferramentas computacionais do Linked Data Stack (van NUFFELEN, 2014), as quais são utilizadas e implementam os processos incrementais da classificação 5-Estrelas; e os conjuntos de dados primários utilizados. Ressalta-se que, metodologicamente, os passos que constituem o *Workflow* proposto são inspirados em um subconjunto de atividades do ciclo de vida Linked Data Lifecycle e tecnicamente suportado pelas ferramentas da Linked Data Stack.

3.1 Linked Data Lifecycle

O procedimento metodológico norteador para o desenvolvimento do *Workflow* proposto baseia-se no Linked Data Lifecycle, um ciclo de vida derivado das práticas do projeto *LOD2 - Creating knowledge out of Interlinked Data* (AUER, 2014). Tal projeto é um empreendimento conjunto de grupos de pesquisa de vanguarda na evolução das metodologias e tecnologias voltadas ao tratamento de Dados Abertos Conectados. O Linked Data Lifecycle consiste em oito atividades, as quais são executadas conforme os requisitos de publicação de dados conectados. As referidas atividades são: Extração (*Extraction*); Armazenamento / Consulta (*Storage / Querying*); Revisão Manual / Autoria (*Manual Revision / Authoring*); Interligação / Fusão (*Interlinking / Fusion*); Classificação / Enriquecimento (*Classification / Enrichment*); Análise de Qualidade (*Quality Analysis*); Evolução / Reparo (*Evolution / Repair*); e Busca / Navegação / Exploração (*Search / Browsing / Exploration*). Conforme os requisitos deste trabalho, o subconjunto de atividades empregado (Figura 2) é formado pelas atividades:

Figura 2: Representação do Linked Data Lifecycle.
Fonte: traduzido a partir de (AUER, 2014).

- **Extração (*Extraction*)** – visa extrair os dados de suas fontes originais. Salienta-se que os dados podem estar codificados em formatos estruturados e não estruturados. A extração deve atender aos requisitos do tratamento dos diferentes formatos em que os dados são armazenados originalmente, transcrevendo os dados extraídos para um formato padrão de processamento.
- **Classificação/Enriquecimento (*Classification/Enrichment*)** – visa utilizar as ontologias ou os vocabulários para modelar e representar os dados primários em um formato compatível para suportar as atividades de recuperação de dados nos ambientes da Web de Dados.
- **Armazenamento/Consulta (*Storage/Querying*)** – prima pela utilização de soluções computacionais para armazenar e recuperar dados em atividades. Na Web de Dados, aconselha-se que os dados estejam armazenados de acordo com o padrão RDF (*Resource Description Framework*).
- **Busca/Navegação/Exploração (*Search/Browsing/Exploration*)** – objetiva empregar soluções computacionais para consultar e/ou explorar os Dados Abertos Conectados, de acordo com os objetivos traçados por um usuário.

3.2 Linked Data Stack

Em um *Workflow*, as ferramentas computacionais são recursos importantes para a execução das atividades e o alcance do objetivo traçado. Considerando um *Workflow* para a preservação de Dados Abertos Conectados, as atividades do Linked Data Lifecycle encontram alicerce em um conjunto de ferramentas computacionais *open-source*, o pacote LOD2 Stack (AUER *et al.*, 2012). Do referido pacote, dentre as ferramentas disponibilizadas para extração, transformação e disponibilização de recursos de dados, neste trabalho são empregadas:

- **Sparqlify** – é uma ferramenta disponibilizada para ser executada na plataforma Linux (SPARQLIFY, 2017). Como característica principal, esta ferramenta realiza a conversão de um arquivo CSV (*Comma-Separated Values*) para um arquivo com recursos RDF, conforme um arquivo de configuração declarado segundo a linguagem SML (*Sparqlification Mapping Language*).
- **OpenLink Virtuoso** - é um Sistema Gerenciador de Banco de Dados (SGBD) que apresenta funcionalidades de manipulação de tabelas de dados relacionais, manipulação de dados em grafos, manipulação de conteúdo XML (*eXtensible Markup*

Language), serviços web, implementação de Dados Abertos Conectados e servidor de aplicações web. Devido a essa flexibilidade, o OpenLink Virtuoso é um sistema universal que permite acesso, integração e gerenciamento de dados baseados no modelo RDF na web (OPENLINK, 2017).

3.3 Conjuntos de dados primários

No âmbito da verificação do *Workflow*, são utilizados três conjuntos de dados primários do domínio da Cientometria, sendo eles:

- **SJR (SCImago Journal & Country Rank)**. O Journal SCImago & Country Rank é um portal que disponibiliza informações cientométricas a partir de dados contidos na base de dados Scopus. Dentre as informações disponibilizadas, está o índice SJR, o qual pode ser utilizado para avaliar a qualidade e a reputação de periódicos científicos (SCOPUS, 2017). Este índice foi coletado no referido portal, em formato XLS (formato compatível com a 2ª Estrela de acordo com a classificação 5-Estrelas), com o período de referência de 2005 a 2016.
- **SNIP (Source Normalized Impact per Paper)**. O índice SNIP é uma métrica que mede o impacto de citação contextual de uma comunicação científica, normalizando a distância interna das citações das comunicações de um periódico perante o universo das citações em uma área de conhecimento (SCOPUS, 2017). Em outras palavras, o SNIP é definido como a razão do impacto bruto de um jornal/revista por publicação e o potencial de citação nas áreas de conhecimento. Isto permite, por exemplo, a avaliação de uma revista em comparação com seus pares com informações mais contextualizadas, fornecendo uma melhor imagem do impacto em determinado domínio. O SNIP também foi coletado nos anos 2015 e 2017. A partir do Portal Journal Metrics, os dados primários são extraídos em formato XLS, com o período de referência de 2005 a 2016.
- **Qualis**. Segundo CAPES (2013), "Qualis é o conjunto de procedimentos utilizados pela CAPES para estratificação da qualidade da produção intelectual dos programas de pós-graduação". O Qualis afere a qualidade de produções científicas a partir da análise da qualidade dos periódicos científicos. Sua classificação compreende oito estratos em ordem decrescente de valor: A1, A2, B1, B2, B3, B4, B5 e C. O índice Qualis foi coletado ao longo dos últimos 12 anos, a partir do Sistema WebQualis (CAPES, 2013) e da Plataforma Sucupira (SUCUPIRA, 2017). Cabe ressaltar que a preservação do índice Qualis como Dados Abertos Conectados foi discutida em Rautenberg e Burda (2016).

4 DEFININDO O *WORKFLOW*

Conforme ilustrado na Figura 3, para elevar os conjuntos de dados Qualis, SNIP e SJR das 1ª e 2ª Estrelas ao patamar da 5ª Estrela, segundo a classificação de abertura de dados 5-Estrelas (5-STAR, 2017), o *Workflow* proposto é constituído com os seguintes passos:

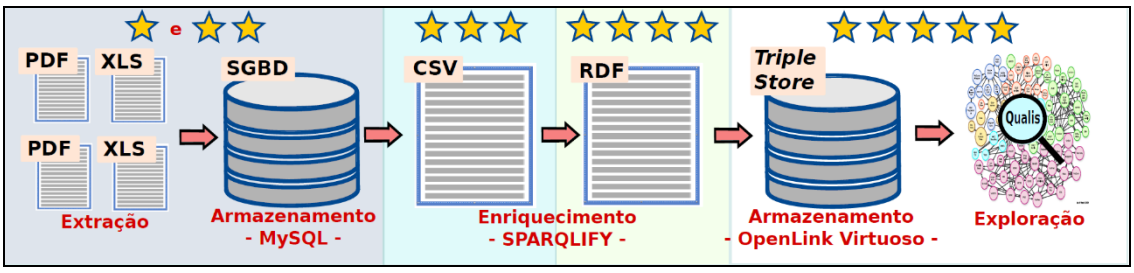

Figura 3: A representação do Workflow proposto.
Fonte: Dados da Pesquisa, 2017.

- **Passo 01 - Extração** – o objetivo deste passo é compatibilizar os dados primários das 1ª e 2ª Estrelas em um formato adequado de processamento ao nível da 2ª Estrela. Ou seja, os arquivos em formato original (PDF – em 1ª Estrela – e XLS em 2ª

Estrela) são convertidos para arquivos de texto passíveis de serem processados computacionalmente.

- **Passo 02 - Armazenamento** – neste estágio, alguns *scripts* de pré-processamento (na linguagem de programação PHP) são empregados para organizar e criticar os dados. Uma vez que os dados foram criticados, tem-se como objetivo realizar o estagiamento desses dados no SGBD Mysql. Tal feita, facilita os processos de manipulação e conversão, minimizando os esforços para inserir os dados primários nas bases de dados de sistemas legados ou representar em outros formatos, por exemplo. Neste ponto, os dados ainda mantêm o patamar da 2ª Estrela.
- **Passo 03 – Enriquecimento** – o propósito deste passo é mapear os dados de um formato proprietário para o formato RDF, transpondo o nível de abertura dos dados da 2ª para a 4ª Estrela. No *Workflow*, a partir de uma base de dados legada, os dados são recuperados e convertidos para arquivos no formato CSV, alcançando temporariamente a 3ª Estrela. Adiante, os dados são mapeados para o formato RDF, de acordo com o modelo de representação discutido em Rautenberg *et al.* (2017). Isso é realizado mediante a utilização da ferramenta Sparqlify. Neste estágio, os dados atingem o patamar da 4ª Estrela.
- **Passo 04 - Armazenamento** – neste passo, visa-se disponibilizar os dados de modo a privilegiar sua reutilização em ambientes da web. Neste sentido, os recursos dos índices Qualis, SJR e SNIP são compartilhados na Web de Dados em um *endpoint* implementado no servidor Open Link Virtuoso. Os dados podem ser relacionados e recuperados via consultas em linguagem SPARQL[1] (DUCHARME, 2013) a partir do endereço http://lod.unicentro.br/sparql. Desta maneira, alcança-se o nível mais elevado na abertura de dados, a 5ª Estrela.
- **Passo 05 – Exploração** – no *Workflow*, este passo tem como objetivo a verificação se a disponibilidade dos dados está correta. Para tanto, consultas na linguagem SPARQL podem ser submetidas ao endpoint http://lod.unicentro.br/sparql, adquirindo informação contextualizada a partir dos recursos de dados dos índices Qualis, SJR e SNIP. Neste ponto, verifica-se o alcance do 5° nível de abertura de dados (a 5ª Estrela).

5 VERIFICAÇÃO DO *WORKFLOW*

Esta seção é organizada em conformidade com os passos definidos para o *Workflow* proposto. A cada subseção, são apresentados os elementos principais, de modo a exemplificar a execução dos passos do *Workflow* ao publicar o histórico do índice SJR como Dados Abertos Conectados. Salienta-se que os mesmos procedimentos são aplicados na publicação dos índices Qualis e SNIP.

5.1 Passo 01 - Extração

Originalmente, os dados abertos dos índices Qualis, SNIP e SJR são acessados em formatos proprietários a partir da web. Neste sentido, considerando a Classificação 5-Estrelas, destaca-se que:

- os dados do índice Qualis capturados do Sistema WebQualis (CAPES, 2013) estavam na 1ª Estrela, no formato PDF;
- os dados dos índices SNIP e SJR são disponibilizados conforme a 2ª Estrela, no formato XLS; e
- da Plataforma Sucupira, o índice Qualis é recuperado no formato XLS.

[1] Acrônimo de Simple Protocol And Rdf Query Language é uma linguagem de consulta para extrair informações de dados baseados em triplas.

Listagem 1: Exemplo de arquivo texto do índice SJR passível de processamento computacional.

```
01 nomeJornal*#*SNIP*#*SJR*#*codigoSubAreaScopus*#*nomeSubAreaScopus*#*issn*#*
   eissn
02 Scientometrics*#*1,319*#*1,154*#*3300*#*General Social
   Sciences*#*01389130*#*15882861
03 Scientometrics*#*1,319*#*1,154*#*3309*#*Library and Information
   Sciences*#*01389130*#*15882861
04 Scientometrics*#*1,319*#*1,154*#*1706*#*Computer Science
   Applications*#*01389130*#*15882861
```

Fonte: Dados da Pesquisa, 2017.

No *Workflow*, o primeiro passo executado padronizou os dados primários para um formato adequado de processamento. Neste sentido, os dados dos índices cientométricos são convertidos para arquivos de texto de modo que seus recursos sejam processados computacionalmente. A

Listagem 1 exemplifica parte do arquivo referente ao índice SJR, no qual os recursos de dados são demarcados pelos caracteres "*#*".

5.2 Passo 02 - Armazenamento

Neste estágio, a partir dos arquivos texto, alguns *scripts* de pré-processamento (na linguagem de programação PHP) foram empregados para organizar e criticar os dados. O passo 02 objetiva realizar o estagiamento dos dados primários no SGBD Mysql. Tal feita, facilita os processos futuros de manipulação e conversão da representação dos dados em outros formatos. Neste ponto, os dados estão no patamar da 2ª Estrela.

Listagem 2: Algoritmo para armazenamento dos dados primários no SGBD MySQL.

```
01 <?php
02  $ano = 2016;
03  $arquivo = fopen ("temp\\txt\\SJR_2016.txt","r");
04  $registro = trim(fgets($arquivo,4096));
05  while (!feof ($arquivo)) {
06   inserirNovasSubAreasScopus($registro);
07   inserirNovosJornais($registro);
08   inserirScoreSJRJornal($registro);
09   $registro = trim(fgets($arquivo,4096));
10  }
11 ?>
```

Fonte: Dados da Pesquisa, 2017.

Na Listagem 2, é apresentado o algoritmo de manipulação dos dados primários do índice SJR para o ano 2016. Em resumo, no algoritmo, a partir do arquivo denominado SJR_2016.txt (linha 3), os dados primários são capturados e inseridos no banco de dados. À medida que os dados são manipulados, verifica-se a necessidade da inserção de novas áreas de conhecimento, novos jornais e escores SJR dos jornais (linhas 6 a 8).

5.3 Passo 03 – Enriquecimento

Com os dados armazenados em uma base de dados de estagiamento, o próximo passo do *Workflow* elevou os dados primários às 3ª e 4ª Estrelas. Neste sentido, com uma consulta ao SGBD Mysql, os dados dos históricos dos índices são recuperados e armazenados em um arquivo de extensão CSV (3ª Estrela). De posse do arquivo CSV e de um arquivo de

configuração da ferramenta Sparqlify, os dados são mapeados para o formato RDF. Ao final deste estágio, os dados atingem o patamar da 4ª Estrela.

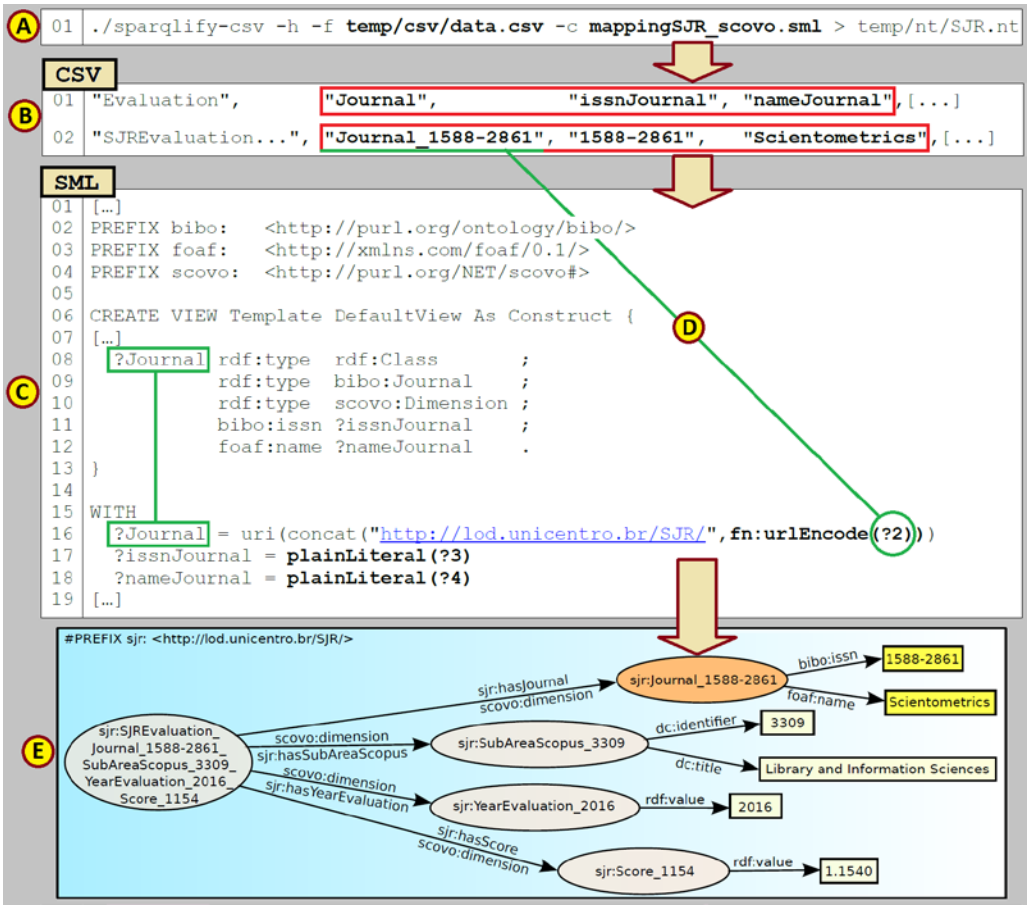

Figura 4: Representação da aplicação da ferramenta Sparqlify.
Fonte: Dados da Pesquisa, 2017.

Para demonstrar este passo, a Figura 4 ilustra a execução da ferramenta Sparqlify nos dados contidos no arquivo CSV, mapeando os dados de acordo com um modelo de representação em RDF (Figura 4.a). Os dados contidos no arquivo CSV (parcialmente exemplificados na Figura 4.b) são mapeados de acordo com um arquivo de configuração (Figura 4.c). Pontualmente, o trecho do arquivo de configuração destacado mapeia os dados relativos aos periódicos de publicação. Destaca-se que as linhas 6-12 modelam o recurso ?Journal e as linhas 16-18 definem quais dados serão mapeados. Por exemplo, a linha 16 mapeia a coluna 2 do arquivo CSV (Figura 4.d) à definição da URI de ?Journal. Ao final da execução, todos os dados contidos no arquivo CSV serão transcritos para um arquivo de recursos RDF (representados na Figura 4.e), conforme o modelo discutido por Rautenberg *et al.* (2017).

5.4 Passo 04 - Armazenamento

Conforme a Listagem 3, de posse dos recursos RDF produzidos no passo anterior, neste passo um *script* (savingIntoVirtuoso.sh) é executado para armazenar tais recursos no servidor Open Link Virtuoso. Para a execução do *script*, dois parâmetros devem ser definidos: a) o nome do arquivo em que os recursos RDF são estagiados (temp/nt/SRJ.nt); e b) o nome do grafo em que os recursos serão publicados no servidor Open Link Virtuoso (http://lod.unicentro.br/SJR/).

Listagem 3: Comando para armazenamento dos recursos de dados do índice SJR na Web de Dados.

```
01 ./savingIntoVirtuoso.sh temp/nt/SJR.nt http://lod.unicentro.br/SJR/
```

Fonte: Dados da Pesquisa, 2017.

Salienta-se que o *script* apresentado anteriormente também foi utilizado para publicar, além do SJR, os índices Qualis e SNIP. Na Tabela 1 são sumarizados os recursos de dados publicados, ano a ano, sendo disponibilizados com Dados Abertos Conectados na Web de Dados: 865.281 avaliações Qualis; 596.465 avaliações SNIP; e 569.732 avaliações SJR.

Tabela 1: Dados primários compartilhados como Dados Abertos Conectados.

ANO	# AVALIAÇÕES QUALIS	# AVALIAÇÕES SNIP	# AVALIAÇÕES SJR
2005	35.020:	34.253	27.977
2006	35.020	36.342	29.570
2007	35.020	38.628	31.226
2008	54.233	41.184	32.965
2009	54.233	44.571	35.316
2010	54.233	48.484	37.997
2011	107.429	53.286	56.410
2012	107.429	56.195	59.578
2013	107.429	58.291	61.955
2014	108.622	59.888	63.629
2015	44.463	62.161	65.947
2016	122.150	63.182	66.732
TOTAL	**865.281**	**596.465**	**569.732**

Fonte: Dados da Pesquisa, 2017.

Uma vez publicados, os dados primários podem ser relacionados, recuperados e explorados a partir da Web de Dados. Desta maneira, alcança-se o nível mais elevado na abertura de dados, a 5ª Estrela. Neste sentido, a próxima subseção exemplifica como relacionar os Dados Abertos Conectados dos índices Qualis, SJR e SNIP.

5.5 Passo 05 - Exploração

Para verificar a funcionalidade do *Workflow*, este passo explorou os Dados Abertos Conectados referentes aos índices cientométricos, os quais foram disponibilizados no endpoint <http://lod.unicentro.br/sparql>. Para tanto, uma consulta na linguagem SPARQL (Listagem 4) foi submetida ao referido endpoint, adquirindo informação contextualizada a partir dos recursos de dados dos índices Qualis, SJR e SNIP. Exemplarmente, a referida consulta relaciona os índices cientométricos associados a um periódico. Neste caso, recupera-se os fatores e as áreas de conhecimento para os índices Qualis, SJR e SNIP do periódico "Scientometrics" no ano 2016.

Listagem 4: Consulta SPARQL do periódico Scintometrics e seus índices cientométricos no ano 2016.

```
01  PREFIX qualis: <http://lod.unicentro.br/QualisBrasil/>
02  PREFIX sjr: <http://lod.unicentro.br/SJR/>
03  PREFIX snip: <http://lod.unicentro.br/SNIP/>
04  PREFIX dc: <http://purl.org/dc/elements/1.1/>
05  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
06  PREFIX foaf: <http://xmlns.com/foaf/0.1/>
07  PREFIX bibo: <http://purl.org/ontology/bibo/>
08
09  SELECT ?issn ?name ?year ?nameAreaCNPq ?qualis ?nameAreaScopus ?snip ?sjr WHERE {
10    ?EvaluationQualis     qualis:hasJournal         ?JournalQualis .
11    ?EvaluationQualis     qualis:hasKnowledgeField  ?KnowledgeField .
12    ?EvaluationQualis     qualis:hasScore           ?ScoreQualis .
13    ?EvaluationQualis     qualis:hasYearEvaluation  ?YearEvaluationQualis.
14    ?JournalQualis        bibo:issn                 ?issn .
15    ?JournalQualis        foaf:name                 ?name .
16    ?KnowledgeField       dc:title                  ?nameAreaCNPq .
17    ?ScoreQualis          rdf:value                 ?qualis.
18    ?YearEvaluationQualis rdf:value                 ?year .
19
20    ?EvaluationSJR      sjr:hasJournal         ?JournalSJR .
21    ?EvaluationSJR      sjr:hasScore           ?ScoreSJR .
22    ?EvaluationSJR      sjr:hasYearEvaluation  ?YearEvaluationSJR.
23    ?JournalSJR         bibo:issn              ?issn .
24    ?YearEvaluationSJR  rdf:value              ?year .
25    ?ScoreSJR           rdf:value              ?sjr.
26
27    ?EvaluationSNIP      snip:hasJournal         ?JournalSNIP .
28    ?EvaluationSNIP      snip:hasScore           ?ScoreSNIP .
29    ?EvaluationSNIP      snip:hasYearEvaluation  ?YearEvaluationSNIP.
30    ?EvaluationSNIP      snip:hasSubAreaScopus   ?SubAreaScopus .
31    ?SubAreaScopus       dc:title                ?nameAreaScopus .
32    ?JournalSNIP         bibo:issn               ?issn .
33    ?YearEvaluationSNIP  rdf:value               ?year .
34    ?ScoreSNIP           rdf:value               ?snip.
35    FILTER (?year = "2016" && ?issn = "1588-2861")
36  }
```

Fonte: Dados da Pesquisa, 2017.

A Figura 5ilustra a submissão da consulta desenvolvida, sendo apresentados: (a) o endereço do endpoint (<http://lod.unicentro.br/sparql>); e (b) a interface para consulta dos recursos. Já o Quadro 1 aponta alguns recursos recuperados pela consulta.

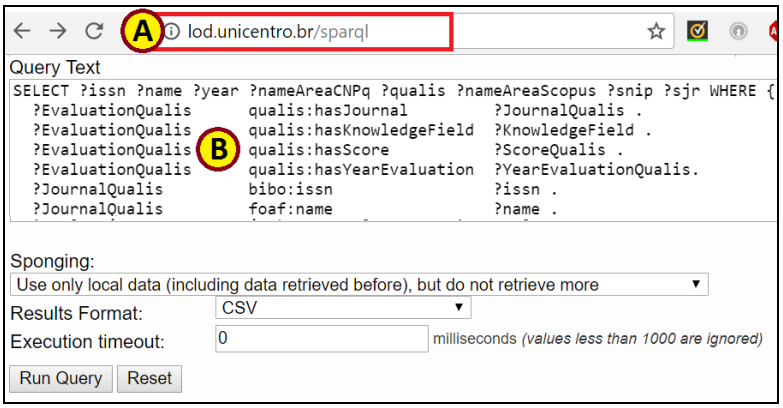

Figura 5: Representação do endpoint http://lod.unicentro.br.
Fonte: Dados da Pesquisa, 2017.

Quadro 1: Resultado parcial do processamento da consulta da Listagem 4.

?issn	?name	?year	?name AreaCNPq	?qualis	?name AreaScopus	?snip	?sjr
[...]	[...]	[...]	[...]	[...]	[...]		[...]
1588-2861	Scientometrics	2016	Ciência da Computação	A1	Computer Science Applications	1.3190	1.1540
[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]
1588-2861	Scientometrics	2016	Administração Pública e de Empresas, Ciências Contábeis e Turismo	A1	Library and Information Sciences	1.3190	1.1540
[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]

Fonte: Dados da Pesquisa, 2017.

Com a consulta desenvolvida, exemplifica-se a possibilidade do reuso e dos cruzamentos de recursos de dados cientométricos disponibilizados pelo *Workflow* proposto. Neste sentido, confirmou-se o alcance do quinto nível de abertura de dados (a 5ª Estrela) para os índices publicados como Dados Abertos Conectados.

6 CONSIDERAÇÕES FINAIS

Este artigo relata o desenvolvimento de um *Workflow* para publicação de dados cientométricos na Web de Dados. Os passos do *Workflow* proposto baseiam-se no ciclo de vida de dados conectados denominado Linked Data Lifecycle (AUER, 2014) e nas ferramentas computacionais do Linked Data Stack (AUER *et al.*, 2012). Ao executar o *Workflow*, os índices cientométricos Qualis, SJR e SNIP são codificados como Dados Abertos Conectados, elevando os níveis de abertura de dados ao quinto nível, de acordo com a classificação 5-Estrelas (5-STAR, 2017).

Ao executar o *Workflow* proposto, verifica-se sua adequação para publicar dados abertos científicos na Web de Dados de forma padronizada. No âmbito da Ciência da Informação, admite-se que o estabelecimento do *Workflow* proposto colabora, principalmente, à preservação digital de dados científicos primários, à reusabilidade de recursos de dados e à reprodutibilidade de resultados em pesquisas.

Neste sentido, a preservação digital de dados científicos primários é alcançada ao atender os requisitos de planejamento e execução do *Workflow* e da evolução dos dados. Ou seja, com a especificação de um conjunto de passos necessários para a transformação dos dados primários em Dados Abertos Conectados, um processo padronizado é planejado cronologicamente. Em um ambiente controlado (com as ferramentas computacionais, os arquivos de configuração e os conjuntos de dados primários são disponibilizados), os passos do *Workflow* para extração, transformação e disponibilização de recursos de dados são orquestrados e executados. E com a utilização do *Workflow* ao longo do tempo, à medida que novos dados primários são gerados/refinados, os dados científicos primários são publicados em conformidade ao quinto nível da classificação 5-Estrelas.

Ao publicar os dados científicos como Dados Abertos Conectados na Web de Dados, aumenta-se as chances do reuso desses recursos em escala global. Em outras palavras, elevando os dados primários científicos ao quinto nível de abertura de dados considerando a classificação 5-Estrelas, a recuperação, a manipulação automatizada de dados, a integração dos dados entre fontes distintas e a interoperabilidade de dados são atividades facilitadas. Por isso, considerando o compartilhamento dos índices Qualis, SJR e SNIP e a possibilidade de reutilizar estes recursos, minimiza-se os esforços de coleta de dados e potencializa-se a consequente reprodutibilidade de resultados em pesquisas científicas no domínio da Ciência da Informação.

Diante as assertivas anteriores, como trabalho futuro, vislumbra-se o uso do *Workflow* proposto como base para:

- atuar na preservação dos recursos de dados dos índices Qualis, SJR e SNIP disponibilizados ao longo do tempo; e
- modelar demais instrumentos de gestão de outros conjuntos de dados cientométricos no escopo das universidades brasileiras, fomentando o desenvolvimento de um "Modelo para Compartilhamento de Informações sobre Pesquisas baseado em *Linked Open Data* para Estudos Cientométricos".

AGRADECIMENTOS

O autor principal agradece à Fundação Araucária pelo suporte financeiro (Projeto n° 601/2014 - Modelo para Compartilhamento de Informações sobre Pesquisas baseado em *Linked Open Data* para Estudos Cientométricos).

REFERÊNCIAS

5-STAR. **5-Star OPEN DATA**. Disponível em: <http://5stardata.info/en>. Acesso em: 15 nov. 2017.

ARELLANO, M. Á. M. **Critérios para a Preservação Digital da Informação Científica**. 2008. 354f. Tese (Doutorado em Ciência da Informação) – Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação da Universidade de Brasília, Brasília, 2008.

AUER, S. Introduction to LOD2. In: AUER, S.; BRYL, V.; TRAMP, C (Ed). **Linked Open Data – Creating Knowledge Out of Interlinked Data**. Heidelberg: Springer-Verlag, 2014. 215p.

AUER, S.; BÜHMANN, L.; DIRSCHL, C.; ERLING, O.; HAUSENBLAS, M.; ISELE, R.; LEHMANN, J.; MARTIN, M.; MENDES, P. N.; van NUFFELEN, B.; STADLER, C.; TRAMP, S.; WILLIAMS, H. Managing the Life-Cycle of Linked Data with the LOD2 Stack. **Lecture Notes in Computer Science**, v. 7650, p 1-16, 2012.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR. **Sistema WebQualis: Portal Capes**. Disponível em: <http://qualis.capes.gov.br/webqualis/principal.seam>. Acesso em: 25 ago. 2013.

CUNHA, M. B. da; CAVALCANTI, C. R. de O. **Dicionário de biblioteconomia e arquivologia**. Brasília: Briquet de Lemos, 2008. 451p.

DUCHARME, B. **Learning SPARQL querying and updating with SPARQL 1.1**. Sebastopol: O'Reilly Media, 2013. 386p.

FERNANDES, L. **Sistemas de gestão documental e workflow no contexto da gestão da qualidade**, 2012. 163f. Dissertação (Mestrado em Ciência da Informação) - Faculdade de Engenharia da Universidade do Porto, Porto, 2012.

HEATH, T.; BIZER, C. **Linked Data Evolving the Web into a Global Data Space**. Amsterdam: Morgan & Claypool, 2011. 136p.

HOLLINGSWORTH, D. **The Workflow Reference Model.** Hampshire: Workflow Management Coalition, 1995. Disponível em: <http://www.wfmc.org/docs/tc003v11.pdf>. Acesso em: 15 de nov. 2017.

OPEN KNOWLEDGE INTERNATIONAL. **Open Data Handbook - o que são Dados Abertos?** Disponível em: <http://opendatahandbook.org/guide/pt_BR/what-is-open-data/>. Acesso em: 14 jun 2017.

OPENLINK. **OpenLink Virtuoso**. Disponível em: <http://virtuoso.openlinksw.com/>. Acesso em: 15 nov. 2017.

RAUTENBERG, S.; BURDA, A. C. Linked Open Data para Cientometria: Compartilhando e Mantendo o índice Qualis na Web de Dados In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais...** São Paulo: USP, 2016. p. A34.

RAUTENBERG, S.; SOUZA, L.; HAUAGGE, J.; HILD, T.; MICHELON, G.; BURDA, A. Representando índices cientométricos como Dados Abertos Conectados. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18, 2017, Marília. **Anais ...** Marília-SP: PPGCI, UNESP, 2017.

SCOPUS. **Journal Metrics - Scopus.com.** Disponível em: <https://www.journalmetrics.com/>. Acesso em: 15 nov. 2017.

SPARQLIFY. **Sparqlify - Agile Knowledge Engineering and Semantic Web (AKSW)**. Disponível em: <http://aksw.org/Projects/Sparqlify.html>. Acesso em: 15 nov. 2017.

SUCUPIRA. **Plataforma Sucupira**. Disponível em: <https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf>. Acesso em: 3 abr. 2017.

van NUFFELEN, B.; JANEV, V.; MARTIN, M.; MIJOVIC, V.; TRAMP, S. Supporting the linked data life cycle using an integrated tool stack. in: AUER, S.; BRYL, V.; TRAMP, S. (Eds.). **Linked Open Data – Creating Knowledge Out of Interlinked Data**. Springer Verlag, 2014.

W3C. **Data on the Web best practices: W3C recommendation.** Disponível em: <https://www.w3.org/TR/2017/REC-dwbp-20170131/>. Acesso em: 22 mar. 2017.

Editores do artigo: Enrique Muriel-Torrado, Edgar Bisset Alvarez, Camila Barros.