

Curadoria Digital de dados e Web de Dados: mantendo Dados Abertos Conectados para estudos bibliométricos e cientométricos

Sandro Rautenberg

Doutor; Universidade Estadual do Centro-Oeste, Guarapuava, PR, Brasil;
srautenberg@unicentro.br

Tony Alexander Hild

Mestre; Universidade Estadual do Centro-Oeste, Guarapuava, PR, Brasil;
thild@unicentro.br

Lucélia de Souza

Doutora; Universidade Estadual do Centro-Oeste, Guarapuava, PR, Brasil;
lucelia@unicentro.br

Resumo: Discorre sobre os primeiros passos em direção à Curadoria Digital de índices cientométricos como Dados Abertos Conectados na Web de Dados. Por se tratar de um relato de experiência, com sua base constitutiva permeando a interdisciplinaridade, o procedimento metodológico é baseado nos ciclos de vida da Curadoria Digital proposto pelo *Digital Curation Centre* e do *Linked Data Lifecycle* proposto pelo grupo de pesquisa *Agile Knowledge Engineering and Semantic*. Como resultado, implementa-se o endpoint <<http://lod.unicentro.br>>, custodiando os recursos de dados do histórico de três índices cientométricos (Qualis, SJR e SNIP, no período 2005-2016). Diante disso, os referidos índices estão disponibilizados na Web de Dados, privilegiando o acesso, o reuso, a interoperabilidade e a processabilidade de seus recursos para com outras pesquisas bibliométricas/cientométricas. Conclui-se que, ao custodiar três conjuntos de índices cientométricos como Dados Abertos Conectados, o endpoint <<http://lod.unicentro.br>> permite: **i)** a encontrabilidade dos recursos na Web de Dados; **ii)** a navegabilidade entre os elementos distribuídos na web; **iii)** a exploração informacional do relacionamento entre os recursos disponibilizados; e **iv)** a replicação de resultados em estudos científicos ao longo do tempo.

Palavras-chave: Curadoria Digital. Web Semântica. Cientometria. Gestão da Informação. Bases de Dados Científicos.

1 Introdução

Com os avanços e a massificação da utilização de recursos das Tecnologias da Informação e Comunicação, principalmente da Internet, é notório que a humanidade tem inovado e maximizado as formas de produção e de emprego dos dados, informações e conhecimento. As plataformas digitais empoderaram o indivíduo como agente principal na socialização de conteúdo, vídeos, e-mails, *posts*, entre outras formas de registro (AALST, 2014). Neste contexto, em uma realidade enraizada na liberdade de distribuição e de uso de recursos digitais, a Web de Dados desponta como uma plataforma global que permite a publicação e o compartilhamento de Dados Abertos Conectados (AUER, 2014). Em poucas palavras, a Web de Dados possibilita a exploração de recursos de dados nos mais variados domínios (W3C, 2018a), conferindo aos dados disponibilizados os benefícios de (W3C, 2018b): **a) reuso** – aumenta as chances de reutilização de dados publicados por consumidores, independentemente do domínio de aplicação; **b) compreensão** – com a publicação de conjuntos de dados e seus respectivos metadados, os seres humanos tem melhor entendimento das natureza, estrutura e semântica dos dados compartilhados; **c) interligação** – por publicar os dados em consonância com os protocolos da Internet, possibilita-se a criação de relações (*hiperlinks*) entre recursos de dados (conjuntos de dados e itens de dados); **d) descoberta** – mediante os *hiperlinks*, autonomamente, os computadores descobrem os conjuntos de dados e navegam entre os recursos digitais, aferindo conhecimento relacional entre os itens compartilhados; **e) confiança** – com a publicação de metadados de proveniência, os publicadores expressam como os conjuntos de recursos digitais são custodiados ao longo do tempo; **f) acesso** – utilizando os protocolos de acesso da Internet, promove-se a encontrabilidade e a exploração de recursos digitais pelos agentes (humanos e de *software*); **g) interoperabilidade** – permite que os recursos digitais sejam automaticamente convertidos em formatos abertos diversos (CSV¹, JSON², RDF³, XML⁴, entre outros); **h) processabilidade** – adicionalmente à interoperabilidade, a processabilidade permite que as aplicações computacionais manipulem os recursos digitais contidos na Web de Dados de forma automática.

Ressalta-se que o acolhimento desses benefícios ocorre com a implantação dos *endpoints* na estrutura global da Web de Dados. Um *endpoint* indica um endereço de Internet de acesso a um serviço da web de onde, mediante o uso de um protocolo específico de consulta, pode-se recuperar os recursos digitais em um formato específico (W3C, 2018c). Em outras palavras, um *endpoint* permite que os agentes (humanos ou de *software*) consultem as bases de conhecimento disponibilizadas na Web de Dados. Geralmente, os resultados das consultas são retornados em formatos abertos legíveis por humanos e/ou processáveis em máquina (SEMANTIC WEB, 2018).

Para este trabalho, parte-se da premissa que a implantação de um *endpoint* envolve os preceitos da Curadoria Digital (DIGITAL CURATION CENTER, 2018a) e do Ciclo de Vida de Dados Conectados (AUER, 2014). Ou seja, as atividades da Curadoria Digital são realizadas para manter, preservar e agregar valor (DIGITAL CURATION CENTER, 2018b) aos Dados Abertos Conectados.

Conforme Abbott (2018), a Curadoria Digital é envolta de ações direcionadas à preservação de recursos digitais, conservando as fontes informacionais e as interfaces de acesso aos atuais e novos consumidores de dados. Portanto, alinhando esta perspectiva aos preceitos dos Dados Abertos Conectados, colabora-se principalmente: no acesso perene a recursos digitais confiáveis, primando pelos requisitos de qualidade e proveniência de dados; na minimização dos esforços para obter os recursos digitais abertamente compartilhados, considerando os custos financeiros e computacionais envolvidos; e na cooperação entre os publicadores e consumidores de dados, ao subsidiar a reutilização dos recursos digitais.

Adicionalmente, admite-se que a Web de Dados é uma infraestrutura de vanguarda (AUER, 2014), potencializando a socialização de dados em pesquisas científicas. Isso enseja a necessidade de custodiar os recursos digitais nos *endpoints*, privilegiando: a encontrabilidade de Dados Abertos Conectados na Internet; a navegabilidade entre os elementos compartilhados; a exploração informacional da relação entre os recursos custodiados; e principalmente, a replicação de estudos científicos ao longo do tempo.

Neste sentido, pressupõe-se que pesquisas científicas se beneficiam da Web de Dados, ao mitigar os esforços na coleta de dados primários. Como um relato de experiência, inicialmente apresentado no 6º Encontro Brasileiro de Bibliometria e Cientometria (6º EBBC) (RAUTENBERG; HILD; SOUZA, 2018), este artigo melhor detalha os esforços despendidos ao manter o *endpoint* <<http://lod.unicentro.br>> como uma fonte informacional importante à Ciência da Informação. Pontualmente, no referido *endpoint*, são compartilhados os recursos digitais dos históricos dos índices Qualis, *SCImago Journal & Country Rank* (SJR) e *Source Normalized Impact per Paper* (SNIP).

Para melhor apresentar o *endpoint* <<http://lod.unicentro.br>>, além dessa seção introdutória, este artigo compreende discussões sobre: **a) Fundamentação Teórica** – disserta sobre os conceitos e ciclos de vida dos Dados Conectados e da Curadoria Digital; **b) Materiais e Métodos** – enumera as fontes informacionais originais dos índices cientométricos e o processo interdisciplinar de Curadoria Digital dos Dados Abertos Conectados custodiados; **c) Endpoint** <<http://lod.unicentro.br>>- esclarece a forma de acesso aos recursos custodiados no *endpoint*; **d) Estudos de caso** – exemplifica as consultas em SPARQL⁵ para recuperar os recursos de dados dos índices Qualis, SJR e SNIP; e **e) Considerações Finais** – pontua as assertivas a respeito dos esforços despendidos e dos trabalhos futuros.

2 Fundamentação teórica

A base constitutiva deste trabalho abarca elementos da Ciência da Computação e da Ciência da Informação no tocante à Curadoria Digital de recursos de dados cientométricos publicados na Web de Dados. Para melhor entender a interdisciplinaridade subjacente, nesta seção são apresentados os conceitos e ciclos de vida dos Dados Abertos Conectados e da Curadoria Digital.

O entendimento do que são Dados Abertos Conectados é configurado por duas questões norteadoras: O que são Dados Abertos? E o que são Dados Conectados? Resumidamente, relaciona-se aos Dados Abertos as licenças que permitem o livre uso desses recursos por pessoas ou aplicações computacionais, definindo as regras pertinentes de distribuição e (re)utilização (LINKED DATA,

2018). Alguns exemplos de regras podem ser: a necessidade de citação da fonte original, o uso não comercial dos dados, a não derivação ou não adaptação dos dados, o compartilhamento dos dados em mesmo formato, dentre outros. Por sua vez, os Dados Conectados referem-se aos dados representados mediante os protocolos de acesso da Internet e que usam a Web de Dados como infraestrutura para promover sua publicidade (BIZER; HEATH; BERNES-LEE, 2009). Isto é possível pela utilização de um modelo para relacionar dados de diversas origens. Sob este prisma, tem-se o *Framework* de Descrição de Recursos (*Resource Description Framework* - RDF), uma linguagem padrão para conectar dados na web. Resumidamente, o RDF descreve os recursos em três partes (tripla no formato sujeito – predicado – objeto), relacionando um sujeito a um objeto por meio de um predicado. Como exemplo, a Figura 1 evidencia uma tripla RDF no domínio deste trabalho, na qual um sujeito identificado por “qualis:Journal_1808-5245” tem um predicado (dc:title) que aponta ao nome “Em Questão”.

Figura 1 - Representação de uma tripla RDF



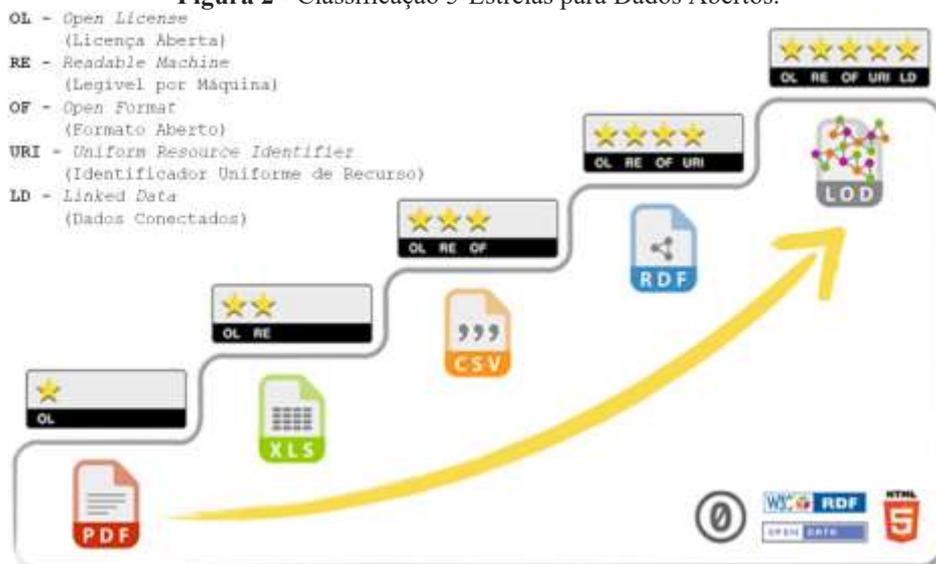
Fonte: Elaborada pelos autores (2018).

Ressalta-se que um conjunto de triplas RDF restrito a um assunto gera uma grande estrutura de ligações, formando um grafo RDF. Neste sentido, o exemplo expresso na Figura 1 faz parte de um grafo, denominado <<http://lod.unicentro.br/QualisBrasil/>>, o qual é discutido na Seção 3.

Constitutivamente, a união das abordagens Dados Abertos e Dados Conectados dão origem aos Dados Abertos Conectados. Neste sentido, os Dados Abertos, Dados Conectados e Dados Abertos Conectados tem sua classificação estabelecida conforme seu nível de acesso e de ligação a outros dados. Essa classificação é representada na Figura 2, a qual é denominada 5-Estrelas (5-Star). Proposta por Tim Bernes-Lee, a referida classificação é organizada incrementalmente como segue (5-STAR, 2018): **1ª Estrela** – é atribuída aos dados abertos que são publicados sob uma licença aberta num formato

proprietário (*Open License* - OL). Os dados neste nível de abertura são manipulados (lidos, visualizados ou impressos) por *softwares*, geralmente, proprietários; **2ª Estrela** – é conferida aos dados abertos legíveis por máquinas (*Readable Machine* - RE), os quais podem ser exportados em outros formatos mediante o uso de *softwares* de edição; **3ª Estrela** – é concedida aos dados abertos que são publicados em formato aberto (*Open Format* - OF), permitindo a manipulação sem a necessidade do uso de *softwares* proprietários; **4ª Estrela** – é designada à utilização de Identificador Uniforme de Recursos (*Uniform Resource Identifier* - URI), semelhante a um endereço de Internet, para rotular os dados, permitindo que outros usuários criem ligações e façam reuso de dados disponibilizados em ambientes web. A partir deste nível, tem-se a possibilidade de publicar os Dados Abertos Conectados; e **5ª Estrela** – é atribuída aos Dados Conectados a outros dados (*Linked Data* – LD). Para tanto, faz-se uso de vocabulários e ontologias conhecidos para permitir a organização, a representação, a navegação entre os recursos de dados e a descoberta de outros conjuntos de dados ou informação relevante. Dessa forma, acrescenta-se valor aos dados, ao facilitar uma contextualização mais ampliada.

Figura 2 - Classificação 5-Estrelas para Dados Abertos.

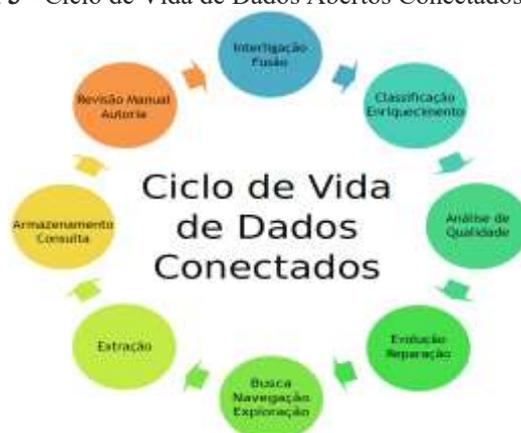


Fonte: Adaptada de 5-STAR (2018).

Cabe ressaltar que, ao publicar os recursos de dados em consonância à 5ª Estrela, fomenta-se uma imensa base informacional denominada Web de Dados. Para tanto, o ato de elevar os recursos de dados dos níveis iniciais de abertura ao

patamar da 5ª Estrela requer a aplicação de procedimentos adequados. Neste sentido, Auer (2014) propõe o Ciclo de Vida de Dados Conectados (do inglês, *Linked Data Lifecycle*). Ilustrado na Figura 3, esse ciclo de vida contempla um conjunto de oito atividades para publicar Dados Conectados: **1) Extração** – os dados não estruturados ou estruturados em diferentes formatos ou provenientes de sistemas legados necessitam ser mapeados para o modelo de dados RDF; **2) Armazenamento/Consulta** – o gerenciamento de dados RDF é realizado por meio do uso de *Triple Stores* como forma de potencializar as tarefas de publicação e consumo de dados; **3) Revisão Manual/Autoria** – as tarefas de editoração de dados são permitidas nesta atividade; **4) Interligação/Fusão** – os dados de uma base são interligados a outros dados de outros conjuntos, ampliando os contextos informacionais; **5) Classificação/Enriquecimento** – seu objetivo é aumentar a expressividade e a riqueza semântica dos dados em relação a um contexto, representando os recursos com ontologias ou vocabulários; **6) Análise de Qualidade** – o tratamento dos aspectos de integridade, precisão, consistência e validade de dados é realizado nesta atividade. De forma geral, verificam-se os quesitos de consistência, concisão, compreensão, disponibilidade e proveniência dos dados; **7) Evolução/Reparação** – uma vez encontradas inconsistências nos dados ou no modelo de representação perante requisitos previamente estabelecidos, ações podem ser tomadas para corrigir as não conformidades; e **8) Busca/Navegação/Exploração** – as técnicas de exploração ou visualização são usadas para manipular os Dados Conectados em diferentes aplicações.

Figura 3 - Ciclo de Vida de Dados Abertos Conectados adotado

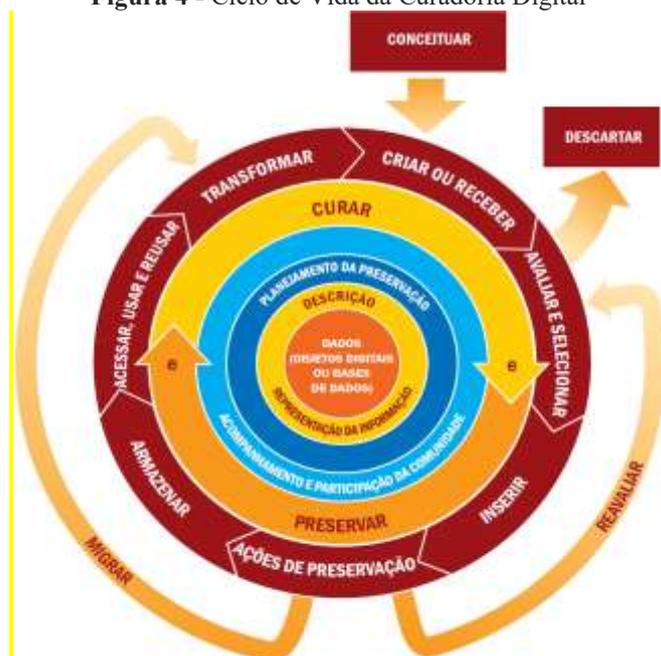


Fonte: Auer (2014, p. 4, tradução dos autores).

Salienta-se que essas atividades são efetivadas de maneira incremental e combinadas conforme os desafios encontrados na publicação dos Dados Abertos Conectados. Neste sentido, algumas questões importantes afloram, por exemplo: Como armazenar os Dados Abertos Conectados na Web de Dados? Como organizar e catalogar os Dados Abertos Conectados e metadados nesse ambiente? Como garantir que os Dados Abertos Conectados estejam adequadamente disponíveis? Interdisciplinarmente, essas questões ensejam a Ciência da Informação, introduzindo competências da Curadoria Digital.

Segundo Siebra, Borba e Miranda (2016), a Curadoria Digital baseia-se em práticas interdisciplinares, combinando aspectos tecnológicos, comunicacionais, gerenciais e cognitivos aplicados à custódia de dados, informação e conhecimento. Considerando o crescimento exponencial da produção de objetos digitais (e-mails, *e-books*, vídeos, imagens, *posts*, registros, dentre outros) mediante o uso de Tecnologias da Informação e Comunicação, esse conceito vem ganhando notoriedade nas Ciência da Computação e Ciência da Informação. Neste contexto, a Curadoria Digital privilegia os estudos quanto à gestão da preservação de recursos digitais nos domínios científicos ou corporativos.

Figura 4 - Ciclo de Vida da Curadoria Digital



Fonte: Digital Curation Centre (2018a, *on-line*, tradução dos autores).

Em suma, a Curadoria Digital se preocupa com todas as atividades envolvidas no emprego de melhores práticas de digitalização e no gerenciamento do ciclo de vida dos objetos digitalizados, assegurando a disponibilidade dos recursos para sua reutilização no futuro (ABBOTT, 2018). Diante disso, a Curadoria Digital (Figura 4) é um conceito vinculado à manutenção, à veracidade e à proveniência, bem como à garantia da qualidade dos dados (ROY; UNDERWOOD; CHANG, 2015), observando a disponibilização perene dos recursos mantidos. Por isso, apropriando-se de um ciclo de vida (DIGITAL CURATION CENTRE, 2018a), considera-se que a Curadoria Digital pode ser aplicada na Web de Dados para: **a) Conceituar** – é a formalização de documentos que definem as orientações, as políticas, os requisitos legais e as ações de criação, representação, captura, limpeza, avaliação e guarda dos dados e metadados em ecossistemas baseados na Web de Dados; **b) Criar ou Receber** – são as ações para criar dados em um ecossistema da Web de Dados. Os metadados decorrentes dessas ações (metadados administrativos, descritivos, estruturais, técnicos e de preservação) também devem ser considerados/mantidos. Na criação ou recebimento de dados, deve-se proceder em consonância às políticas de coleta documentadas na fase Conceituar; **c) Avaliar e Selecionar** – antes de inserir novos dados no ecossistema, deve-se avaliar os dados quanto aos requisitos de qualidade estabelecidos (as orientações, as políticas e os requisitos legais de criação, captura e guarda de dados). Uma vez avaliados, seleciona-se o conjunto íntegro de dados para ser custodiado e preservado; **d) Inserir** – definido o conjunto íntegro de dados, o próximo passo é armazenar os dados no ecossistema da Web de Dados, de acordo os documentos previamente formalizados; **e) Ação de preservação** – realiza-se as ações de garantia da preservação dos dados. As ações de preservação são previamente definidas e são orquestradas para que os dados permaneçam autênticos, confiáveis e usáveis, mantendo perenemente a integridade dos recursos; **f) Armazenar** – ao custodiar os dados e metadados, garante-se que os recursos sejam mantidos seguramente, utilizando tecnologias de armazenamento e representação num ecossistema da Web de Dados; **g) Acesso, uso e reutilização** – possibilitar que os dados sejam facilmente acessíveis pelos usuários. Na Web de Dados, os controles de acesso/autenticação

geralmente não são implementados, visto que as políticas devem privilegiar o livre acesso e a (re)utilização de recursos; **h) Transformar** – em algumas circunstâncias, existe a possibilidade de sumarizar ou derivar novos dados a partir dos recursos armazenados; **i) Descarte** – ocasionalmente, pode-se remover dados (desatualizados, invalidados, ou por orientação legal) conforme as políticas documentadas. Normalmente, os dados são retirados de um ambiente de produção, sendo transferidos para um arquivo morto passível de recuperação. Em outros casos, os dados são definitivamente eliminados, por razões legais que sustentam a destruição segura; **j) Reavaliar** – quando necessário, pode-se reavaliar uma versão mais recente dos dados que anteriormente foram invalidados de acordo com os procedimentos formalizados na fase Conceituar; e **k) Migrar** – em virtude de avanços tecnológicos, deve-se executar ações de migração dos dados para um formato mais atual. Desta forma, preserva-se os dados a longo prazo, mesmo ocorrendo a obsolescência de *hardware* ou de *software* nos ecossistemas da Web de Dados.

Diante o exposto acerca dos Dados Abertos Conectados e da Curadoria Digital, constata-se a afinidade conceitual no tocante à preservação de recursos na Web de Dados, privilegiando a reutilização de dados em contextos científicos e/ou corporativos. Neste sentido, este artigo inter-relaciona os ciclos de vida da Curadoria Digital e dos Dados Conectados na custódia de índices cientométricos em ecossistemas da Web de Dados. Dito isso, a próxima seção é reservada à apresentação dos índices cientométricos custodiados como Dados Abertos Conectados e discussão do procedimento metodológico adotado para tal feita.

3 Materiais e métodos

Nesta seção, são abordados os insumos utilizados para a realização da pesquisa. Como materiais, são apresentadas as fontes originais dos índices cientométricos, sumarizando os recursos digitais abstraídos destas bases e compartilhados como Dados Abertos Conectados. Também se relata o procedimento metodológico adotado, o qual subsidia a elevação dos dados originais (nas 1ª e 2ª Estrelas) ao

nível da 5ª Estrela de abertura de dados e custodia os referidos recursos digitais resultantes.

Neste trabalho são consideradas as bases de dados dos índices cientométricos: **a) Qualis** – coletada nos últimos 12 anos a partir do Sistema WebQualis (WEBQUALIS, 2013) e da Plataforma Sucupira (SUCUPIRA, 2017), conforme inicialmente discutido em Rautenber e Burda (2016); **b) SNIP** - recuperada nos anos 2015 e 2017, no Portal *Journal Metrics* (JOURNAL METRICS, 2017). Originalmente, os dados são extraídos em formato XLS⁶, com o período de referência de 2005 a 2016; e **c) SJR** – coletada do Portal *Journal SCImago & Country Rank* (SJR, 2017) no formato XLS, também considerando o período 2005 a 2016.

Como prática inicial de Curadoria Digital, ressalta-se que os conjuntos de dados enumerados são anualmente recuperados a partir de suas fontes originais e armazenados na base de dados de um sistema legado. Posteriormente, é realizada a socialização dos índices cientométricos na Web de Dados. Devido à complexidade, os detalhes intrínsecos da referida socialização não são apresentados neste artigo. Entretanto, os apontamentos sobre a modelagem dos dados originais como Dados Abertos Conectados e seu processo técnico-metodológico de transformação são apresentados por Rautenber et. al. (2017a) e Rautenber et. al. (2017b), respectivamente.

Tabela 1 - Sumarização dos recursos digitais disponíveis no *endpoint* <<http://lod.unicentro.br>>

ANO REFERÊNCIA	# QUALIS	# SNIP	# 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: Elaborada pelos autores (2018).

Salienta-se que os dados recuperados subsidiam os grafos <<http://lod.unicentro.br/QualisBrasil/>>, <<http://lod.unicentro.br/SJR/>> e

<<http://lod.unicentro.br/SNIP/>> no *endpoint* <<http://lod.unicentro.br/>>. Conforme a Tabela 1, atualmente são custodiados cerca de dois milhões de recursos digitais no referido *endpoint*.

Para custodiar os índices cientométricos como Dados Abertos Conectados, metodologicamente, são aplicadas cinco atividades da Curadoria Digital (DIGITAL CURATION CENTER, 2018a). Representadas na Figura 5, interdisciplinarmente, as essas atividades são mediadas por quatro fases do Ciclo de Vida de Dados Conectados (AUER, 2014), conforme descritas na sequência.



Quadro 1 – Exemplos de recursos recuperados com a consulta SPARQL da Figura 8.

Year	Issn	nameJournal	nameKField	Score
[...]	[...]	[...]	[...]	[...]
2016	“1134-3478”	“Comunicar”	“COMUNICAÇÃO E INFOR...”	A1
2016	“0020-0255”	“Information Sciences”	“COMUNICAÇÃO E INFOR...”	A1
2016	“2318-0889”	“TRANSINFORMAÇÃO”	“COMUNICAÇÃO E INFOR...”	A1
[...]	[...]	[...]	[...]	[...]

Fonte: Elaborado pelos autores (2018).

Ainda de acordo com a Figura 8, linhas 21 a 24, os recursos digitais recuperados são referentes às avaliações da área de “COMUNICAÇÃO E INFORMAÇÃO” que alcançaram o escore “A1” no ano de 2016. O Quadro 1 exemplifica alguns dados recuperados.

O segundo estudo de caso tem como objetivo a recuperação de Dados Abertos Conectados a partir do grafo <<http://lod.unicentro.br/SJR/>>. De acordo com as linhas 21 e 23 da Figura 9, a consulta SPARQL codificada é parametrizada para recuperar as avaliações SJR da área de conhecimento denominada “*Library and Information Sciences*”. São consideradas as avaliações do ano de 2016 e conforme a linha 24, os dados recuperados são classificados pelo fator de impacto do periódico, acessando primeiramente as revistas de maior

escore. O Quadro 2 exemplifica os recursos de dados recuperados neste estudo de caso.

Figura 9 - Exemplo de consulta SPARQL para recuperação de recursos do índice SJR

```

01 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
02 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
03 PREFIX dc: <http://purl.org/dc/elements/1.1/>
04 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
05 PREFIX bibo: <http://purl.org/ontology/bibo/>
06 PREFIX sjr: <http://lod.unicentro.br/SJR/>
07
08 SELECT DISTINCT ?year ?issn ?nameJournal ?nameKField ?score WHERE
09 {
10   ?evaluation rdfs:type sjr:Evaluation .
11   ?evaluation sjr:hasJournal ?journal .
12   ?evaluation sjr:hasYearEvaluation ?yearEvaluation .
13   ?evaluation sjr:hasSubAreaScopus ?knowledgeField .
14   ?evaluation sjr:hasScore ?qualisScore .
15
16   ?journal bibo:issn ?issn .
17   ?journal foaf:name ?nameJournal .
18   ?yearEvaluation rdf:value ?year .
19   ?qualisScore rdf:value ?score .
20   ?knowledgeField dc:title ?nameKField .
21
22   FILTER ((ucase(?nameKField) = "LIBRARY AND INFORMATION
23 SCIENCES")
24           && (?year = "2016"))
} ORDER BY DESC(?score)

```

Fonte: Elaborada pelos autores (2018).

Quadro 2 - Exemplos de recursos recuperados com a consulta SPARQL da Figura 9

Year	Issn	nameJournal	nameKField	Score
[...]	[...]	[...]	[...]	[...]
2016	"1137-5019"	"Cybermetrics"	"Library and Information ..."	"4.7190"
2016	"1047-7047"	"Information Systems Research"	"Library and Information ..."	"4.6840"
2016	"1526-5536"	"Information Systems Research"	"Library and Information ..."	"4.6840"
[...]	[...]	[...]	[...]	[...]

Fonte: Elaborado pelos autores (2018).

Por fim, o terceiro estudo de caso exemplifica como recuperar os Dados Abertos Conectados do grafo <http://lod.unicentro.br/SNIP/>. Assim como no exemplo anterior, são recuperados os recursos de dados sobre as avaliações de periódicos da área de conhecimento “*Library and Information Sciences*”, no ano de 2016 (linhas 21-23 da Figura 10), ordenadas pelo escore. Exemplos dos recursos de dados são listados no Quadro 3.

Figura 10 - Exemplo de consulta SPARQL para recuperação de recursos do índice SNIP

```

01 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
02 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
03 PREFIX dc: <http://purl.org/dc/elements/1.1/>
04 PREFIX foaf: <http://xmlns.com/foaf/0.1/>

```

```

05 PREFIX bibo: <http://purl.org/ontology/bibo/>
06 PREFIX snip: <http://lod.unicentro.br/SNIP/>
07
08 SELECT DISTINCT ?year ?issn ?nameJournal ?nameKField ?score WHERE
09 {
10   ?evaluation      rdf:type                snip:Evaluation .
11   ?evaluation      snip:hasJournal         ?journal .
12   ?evaluation      snip:hasYearEvaluation  ?yearEvaluation .
13   ?evaluation      snip:hasSubAreaScopus   ?knowledgeField .
14   ?evaluation      snip:hasScore          ?qualisScore .
15
16   ?journal          bibo:issn              ?issn .
17   ?journal          foaf:name              ?nameJournal .
18   ?yearEvaluation  rdf:value              ?year .
19   ?qualisScore      rdf:value              ?score .
20   ?knowledgeField  dc:title                ?nameKField .
21
22   FILTER ((ucase(?nameKField) = "LIBRARY AND INFORMATION
23 SCIENCES"))
24           &&( ?year = "2016" )
25 } ORDER BY DESC(?score)

```

Fonte: Elaborada pelos autores (2018).

Quadro 3 - Exemplos de recursos recuperados com a consulta SPARQL da Figura 10

Year	issn	nameJournal	nameKField	score
[...]	[...]	[...]	[...]	[...]
2016	"1137-5019"	"Cybermetrics"	"Library and Information Sciences"	"3.7950"
2016	"0268-4012"	"International Journal of Information Management"	"Library and Information Sciences"	"2.8280"
2016	"0740-624X"	"Government Information Quarterly"	"Library and Information Sciences"	"2.7800"
[...]	[...]	[...]	[...]	[...]

Fonte: Elaborado pelos autores (2018).

Vale destacar que as consultas SPARQL apresentadas nas Figuras 8, 9 e 10 podem ser customizadas e submetidas ao *endpoint* <http://lod.unicentro.br>. Desta forma, tais customizações podem auxiliar a coleta de dados primários em demais pesquisas bibliométricas e/ou cientométricas.

6 Considerações finais

Este artigo discorre sobre um estudo interdisciplinar e aplicado da Curadoria Digital de Dados Abertos Conectados. Resumidamente, relata os esforços dispendidos no desenvolvimento do *endpoint* <http://lod.unicentro.br> como fonte informacional de índices cientométricos na Web de Dados.

Pontualmente, são compartilhados os dados históricos dos índices Qualis,

SJR e SNIP. Salienta-se que tais recursos podem ser livremente recuperados em vários formatos de codificação, tornando o referido *endpoint* um importante instrumento de coleta de dados primários na condução de pesquisas bibliométricas ou cientométricas. Para tal, algumas consultas SPARQLs também são apresentadas, as quais podem ser customizadas em relação a outros estudos. Neste sentido, ao custodiar três conjuntos de Dados Abertos Conectados dos índices cientométricos supracitados, o *endpoint* <<http://lod.unicentro.br>> permite: a encontrabilidade desses recursos na Web de Dados; a navegabilidade entre demais elementos bibliométricos ou cientométricos distribuídos na web; a exploração informacional do relacionamento entre os recursos disponibilizados; e a replicação de resultados em estudos científicos ao longo do tempo.

Dito isso, admite-se que o trabalho contribui a uma discussão interdisciplinar sobre os Dados Abertos Conectados e a Ciência da Informação. Em face disso, no que tange à publicação de Dados Abertos Conectados, este trabalho expressa um exemplo profícuo de novas demandas das competências da Curadoria Digital. Neste sentido, a Curadoria Digital torna-se importante à gestão da preservação de recursos digitais em ecossistemas da Web de Dados, assegurando o acesso e (re)uso dos recursos digitais no futuro.

Num segundo momento, este trabalho também contribui à Ciência da Informação e demais subáreas, elevando os recursos de dados dos índices Qualis, SJR e SNIP ao patamar de abertura de dados da 5ª Estrela. Isso ocorre, principalmente, ao fomentar a disponibilidade, o reuso, a interoperabilidade e a processabilidade desses recursos digitais para com outras pesquisas bibliométricas ou cientométricas. Diante disso, como trabalho futuro, pretende-se atuar continuamente na atualização dos históricos dos índices Qualis, SJR e SNIP no *endpoint* <<http://lod.unicentro.br>>.

Agradecimentos

À Fundação Araucária pelo suporte financeiro ao projeto de pesquisa intitulado *Curadoria Digital e Dados Abertos Conectados: um estudo da preservação de recursos digitais na Web de Dados para estudos cientométricos*.

Referências

5-STAR. **5-Star Open Data**. Disponível em: <<http://5stardata.info/en>>. Acesso em: 04 set. 2018.

AALST, W. van der. Data Scientist: The Engineer of the Future. In: INTEROPERABILITY OF ENTERPRISES SYSTEMS AND APPLICATIONS CONFERENCE (I-ESA'2014), 2014, Albi, France. **Anais...** Heidelberg: Springer, 2014.

ABBOTT, D. **What is Digital Curation?** Disponível em: <http://www.dcc.ac.uk/sites/default/files/documents/resource/briefing-papers/what-is-digital-curation.pdf>. Acesso em: 04 set. 2018.

AUER, Sören. Introduction to lod2. In: **Linked Open Data – Creating Knowledge Out of Interlinked Data**. AUER, S.; BRYL, V.; TRAMP, C (Ed.). Lecture Notes in Computer Science. Springer-Verlag, 2014.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked data: the story so far. **International Journal of Semantic Web and Information Systems**, v. 5, n. 1, p. 1-22, 2009.

DIGITAL CURATION CENTER. **DCC Curation Lifecycle Model**. Disponível em: <http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf>. Acesso em: 04 set. 2018a.

DIGITAL CURATION CENTER. **What is digital curation?** Disponível em: <http://www.dcc.ac.uk/digital-curation/what-digital-curation>. Acesso em: 04 set. 2018b.

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

JOURNAL METRICS. **Journal Metrics - Scopus.com**. Disponível em: <https://www.journalmetrics.com/>. Acesso em: 16 abr. 2017.

LINKED DATA. **Linked Data: design issues**. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 04 set. 2018.

RAUTENBERG, Sandro; BURDA, Alessandra. 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: ENANCIB, 2016.

RAUTENBERG, Sandro; HILD, Tony Alexander.; SOUZA, Lucelia de. Curadoria Digital e Dados Abertos Conectados: o endpoint lod.unicentro.br

como fonte informacional da Web de Dados para estudos bibliométricos e cientométricos. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 6., 2018, Rio de Janeiro. **Anais...** Rio de Janeiro, 2018.

RAUTENBERG, Sandro; SOUZA, Lucelia de.; DALL'AGNOL, Josiane M. H.; HILD, Tony Alexander.; MICHELON, Gisane.; BURDA, Alessandra. Representando índices cientométricos como dados abertos conectados. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 28., 2017, Marília. **Anais...** Marília-SP: PPGCI; UNESP, 2017a.

RAUTENBERG, Sandro; MOTYL, Sandro Kaue; BURDA, Alessandra Cassiana; SILVERIO, Anderson; MOURA, Fabrício Maron de. Dados Abertos Conectados e Gestão do Conhecimento: estudos de caso cientométricos em uma universidade brasileira. **Perspectivas em Ciência da Informação**, v. 22, p. 116-142, 2017b.

ROY, Arnab; UNDERWOOD, Mark; CHANG, Wo. **Big Data Interoperability Framework: Volume 4, Security and Privacy**. Gaithersburg: National Institute of Standards and Technology (NIST), 2015. 75 p. Disponível em:

<https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-4.pdf>.

Acesso em: 28 jul. 2018.

SEMANTIC WEB. **SPARQL Endpoint**. Disponível em:

http://semanticweb.org/wiki/SPARQL_endpoint.html. Acesso em: 04 set. 2018.

SIEBRA, Sandra de Albuquerque; BORBA, Valdeane da Rocha; MIRANDA, Májory Karoline Fernandes de Oliveira. Curadoria Digital: um termo interdisciplinar. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 27., 2016, Salvador. **Anais...** Salvador, BA: UFBA, 2016.

SJR. **Scimago Journal & Country Rank**. Disponível em:

<http://www.scimagojr.com/journalrank.php>. Acesso em: 04 dez. 2017.

SUCUPIRA. **Plataforma Sucupira**. Disponível em:

<https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf>. Acesso em: 03 abr. 2017.

W3C. **Data on the Web best practices: W3C recommendation 31 January 2017**. Disponível em: <<https://www.w3.org/TR/2017/REC-dwbp-20170131/>>. Acesso em: 04 set. 2018b.

W3C. **Linked Data**. Disponível em:

<https://www.w3.org/standards/semanticweb/data>. Acesso em: 04 set. 2017a.

W3C. **Web Services Description Requirements**. Disponível em:

<https://www.w3.org/TR/2002/WD-ws-desc-reqs-20021028/#normDefs>. Acesso em: 09 set. 2018c.

WEBQUALIS. Sistema WebQualis - Portal Capes. Disponível em:
<http://qualis.capes.gov.br/webqualis/principal.seam>. Acesso em: 25 de ago.
2013.

Digital Data Curation and Web of Data: maintaining Linked Open Data for bibliometric and scientometric studies

Abstract: We present the first steps toward the Digital Curation of scientometric indexes as Linked Open Data on the Web of Data. As an interdisciplinary research, the methodological approach is based on the lifecycles of Digital Curation proposed by the Digital Curation Centre and Linked Data Lifecycle proposed by the Agile Knowledge Engineering and Semantic Research Group. As a result, the endpoint <<http://lod.unicentro.br>> is implemented and the historical resource data of three scientometric indexes (Qualis, SJR and SNIP, period 2005-2016) are curated. Therefore, these indexes are available on the Web of Data, promoting the access, reuse, interoperability and processability of digital resources in bibliometric/scientometric researches. We conclude that, by preserving the three datasets as Linked Open Data, the endpoint <<http://lod.unicentro.br>> allows: **i)** the availability of resources on the Web of Data; **ii)** the navigability between the resources distributed on the web; **iii)** the exploration of the relationship between the available elements; and **iv)** the replication of results in scientific studies over time.

Keywords: Digital Curation. Semantic Web. Scientometrics. Information Management. Scientific Datasets.

Recebido: 08/09/2018

Aceito: 05/12/2018



¹ **CSV – Comma Separated Values** é um formato para armazenar dados tabulares em texto, e cada linha do arquivo representa um registro. Cada registro é composto de um ou mais campos separados por vírgula. É usado para representar dados de estruturas simples, diretamente processados por editores de planilha ou de texto.

² **JSON – JavaScript Object Notation** é um formato para troca de dados, de fácil escrita e leitura para pessoas e fácil análise e geração para máquinas. É construído sobre duas estruturas: i) uma coleção de pares nome/valor; e ii) uma lista ordenada de valores. Ambas estruturas são amplamente usadas em linguagens de programação, facilitando seu entendimento e uso.

³ **RDF – Resource Description Framework** é uma linguagem que usa um modelo padrão para troca de dados na web, o qual possui a forma de triplas, realizando a descrição de um recurso em três partes, ligando um sujeito (recurso) a um objeto através de um predicado. Isso gera uma estrutura de ligações que, por sua vez, formam um grafo RDF.

⁴ **XML – Extensible Markup Language** é uma linguagem de marcação extensível que usa um formato padrão para descrição e troca de dados na web. Os dados são organizados de forma hierárquica, permitindo a representação de todos os tipos de estruturas de dados.

⁵ **SPARQL** – *Sparql Protocol And Rdf Query Language* é uma linguagem de consulta da Web de Dados, utilizada para extrair informações de dados que são representados em triplas RDF (DUCHARME, 2013).

⁶ **XLS** – acrônimo de *eXcel Spreadsheet* - formato de planilha eletrônica nativamente utilizado pelo software Excel da Microsoft.