

CURADORIA DIGITAL E DADOS ABERTOS

CONECTADOS: o *endpoint* lod.unicentro.br como fonte informacional da Web de Dados para estudos bibliométricos e cientométricos

Sandro Rautenberg

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

Tony Alexander Hild

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

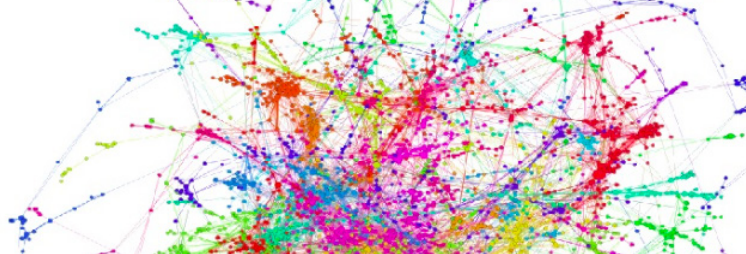
Lucelia de Souza

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

1 INTRODUÇÃO

A Web de Dados é uma plataforma global para compartilhar Dados Abertos Conectados¹, possibilitando a exploração de recursos digitais nos mais variados domínios (W3C, 2017a). Para *World Wide Web Consortium* (W3C, 2017b), o fomento dessa plataforma envolve a adoção de práticas que permitem a obtenção dos seguintes benefícios: **reuso** – aumenta as chances de reutilização de dados pelos consumidores; **compreensão** – propicia aos seres humanos um melhor entendimento das natureza, estrutura e semântica dos dados compartilhados; **interligação** – pos-

1 O entendimento do que são Dados Abertos Conectados relaciona-se com duas definições: a) o que são Dados Abertos; e b) o que são Dados Conectados. Dados Abertos são os dados regidos por licenças que não impedem o (re)uso, desde que obedecidas as regras pertinentes de distribuição (LINKED DATA, 2017). Por sua vez, Dados Conectados se referem a um conjunto de práticas para publicar e interligar dados estruturados usando a web como uma plataforma global (BIZER; HEATH; BERNES-LEE, 2009).

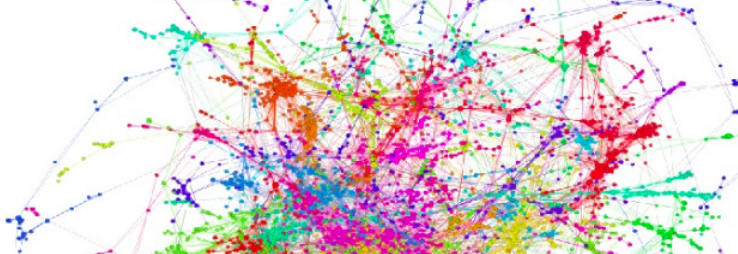


sibilita a criação de relações entre recursos digitais; **descoberta** – habilita os computadores descobrirem de forma autônoma os conjuntos de recursos digitais; **confiança** – certifica que os conjuntos de recursos digitais são curados ao longo do tempo; **acesso** – facilita o acesso dos agentes (humanos e de software) aos recursos atualizados; **interoperabilidade** – permite que os recursos digitais sejam utilizados em aplicações de formatos variados; e **processabilidade** – habilita os computadores a processar os recursos digitais contidos na Web de Dados. Ressalta-se que o acolhimento desses benefícios ocorre com a implantação dos *endpoints*² na estrutura global da Web de Dados.

Na implantação de um *endpoint*, deve-se observar os preceitos da Curadoria Digital (DIGITAL CURATION CENTER, 2017a) e do Ciclo de Vida de Dados Conectados (AUER, 2014). Ou seja, as atividades da Curadoria Digital devem ser realizadas para manter, preservar e agregar valor (DIGITAL CURATION CENTER, 2017b) aos Dados Abertos Conectados.

Conforme Abbott (2017), 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: **a)** no acesso perene a recursos digitais confiáveis, primando pelos requisitos de qualidade e proveniência de dados; **b)** na minimização dos esforços para obter os recursos digitais compartilhados, considerando os custos financeiros e computacionais envolvidos; e **c)** na cooperação entre os publicadores e consumidores de dados, ao subsidiar a reutilização dos recursos digitais disponibilizados.

2 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, se recuperam os recursos digitais em um formato específico (W3C, 2017c). Na Web de Dados, um *endpoint* permite que os agentes (humanos ou de software) consultem uma base de conhecimento através da linguagem SPARQL, por exemplo. Geralmente, os resultados das consultas são retornados em formatos abertos legíveis por humanos e/ou processáveis em máquina (SEMANTIC WEB, 2017).



Diante disto, admite-se que a Web de Dados é uma infraestrutura de vanguarda para socializar dados primários em pesquisas científicas. Isso enseja a necessidade de curar os recursos digitais nos *endpoints*, privilegiando: a encontrabilidade de recursos digitais na Internet; a navegabilidade entre os elementos; a exploração informacional da relação entre os recursos curados; e, principalmente, a replicação de estudos científicos ao longo do tempo.

Ao considerar o domínio da Ciência da Informação, pressupõe-se que suas pesquisas se beneficiam da Web de Dados, ao mitigar os esforços na coleta de dados primários. Este artigo relata os esforços despendidos ao manter o *endpoint* <http://lod.unicentro.br> como uma fonte informacional. Neste sentido, como descrito a seguir, 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).

2 MATERIAIS E MÉTODOS

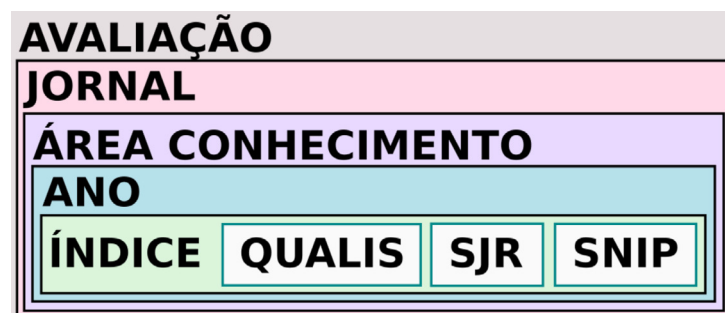
Neste trabalho são consideradas bases de dados dos seguintes índices cientométricos: **a) Qualis** – coletada nos últimos 12 anos no Sistema WebQualis (WEBQUALIS, 2013) e na Plataforma Sucupira (SUCUPIRA, 2017), conforme inicialmente discutido em Rautenberg e Burda (2016); **b) SNIP** – recuperada nos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

Como prática inicial de curadoria, 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 Rautenberg et al. (2017a) e Rautenberg et al. (2017b), respectivamente.

FIGURA 1 - REPRESENTAÇÃO DOS RECURSOS DIGITAIS



Fonte: Dados da pesquisa.

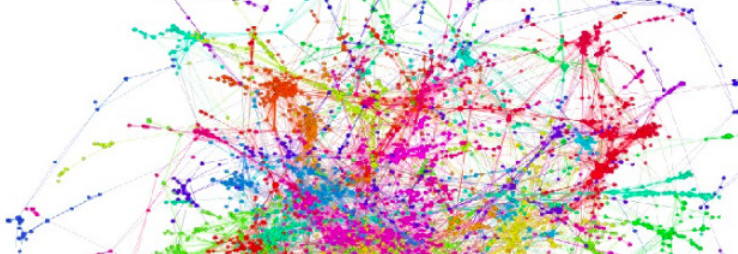
Baseando-se em Rautenberg et al. (2017a), apresenta-se o modelo que formaliza o escore obtido por um jornal dado um índice cientométrico em particular (Figura 1). Estruturalmente, um recurso digital compartilhado compreende uma avaliação de um jornal, atrelando uma área de conhecimento e o ano para o qual o periódico foi avaliado. A essas dimensões de dados, os índices Qualis, SJR e SNIP alcançados são relacionados.

FIGURA 2 - REPRESENTAÇÃO DO PROCEDIMENTO METODOLÓGICO



Fonte: Dados da pesquisa.

Metodologicamente, conforme a Figura 2, neste trabalho são aplicadas cinco atividades da Curadoria Digital (DIGITAL CURATION CENTER, 2017a). Interdisciplinarmente, estas atividades são relacionadas a quatro fases do Ciclo de Vida de Dados Conectados (AUER, 2014), conforme descritas na sequência.



As atividades de Seleção e de Ingestão são realizadas simultaneamente durante a fase de Extração do Ciclo de Vida de Dados Conectados. A Seleção visa recuperar os dados de interesse de um sistema legado para preservação em longo prazo. Com os dados recuperados, a Ingestão transfere os dados para uma área de estágio antes do armazenamento.

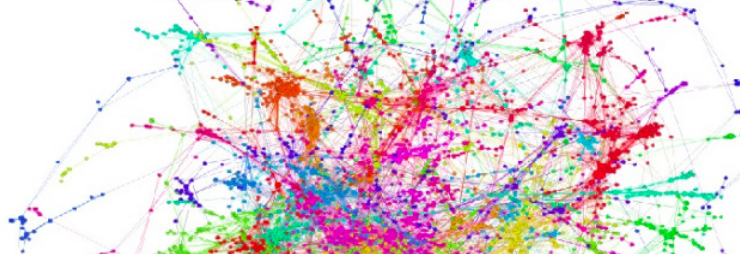
A atividade de Transformação é o ato de criar novos dados a partir dos dados originais, como por exemplo, por meio da migração dos dados originais para outro formato ou do enriquecimento dos dados com semântica. No procedimento metodológico, os dados em formato CSV (*Comma Separated Value*) são convertidos para RDF (*Resource Description Framework*), o modelo padrão de dados da Web de Dados. Simultaneamente, os vocabulários/ontologias (*Bibliografic Ontology* e Dublin Core, por exemplo) são adotados no enriquecimento semântico dos recursos digitais.

Na Curadoria Digital, a atividade de Armazenamento dos recursos digitais deve ser realizada de forma segura, respeitando os padrões relevantes. Neste sentido, para privilegiar o (re)uso de recursos digitais na Web de Dados, os elementos resultantes da atividade de Transformação são armazenados no *endpoint* <http://lod.unicentro.br>.

A atividade de Acesso, Uso e Reuso engloba as ações para garantir que os recursos digitais sejam acessíveis pelos consumidores de dados de forma ininterrupta. No Ciclo de Vida de Dados Conectados, essa atividade é relacionada à fase de Busca, Navegação e Exploração para recuperar e visualizar os recursos digitais compartilhados na Web de Dados.

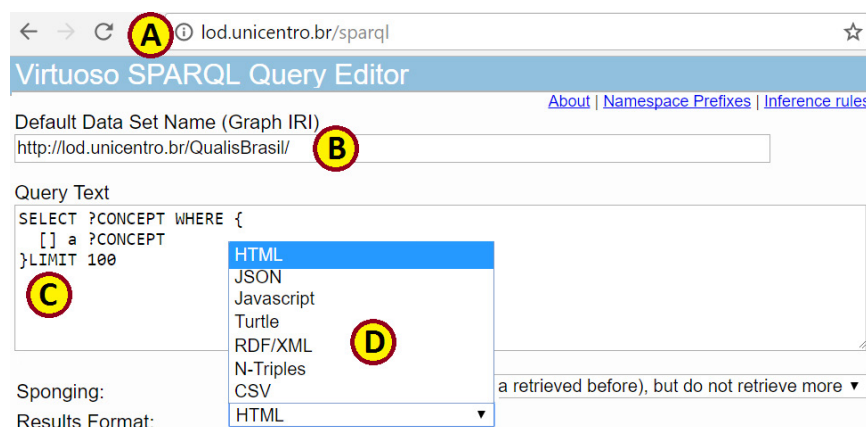
3 RESULTADOS E DISCUSSÃO

Na Figura 3 é representado o *endpoint* lod.unicentro.br, onde: **(A)** apresenta o endereço de Internet disponibilizado aos agentes para acesso dos recursos digitais compartilhados; **(B)** define o conjunto de dados disponibilizado a ser explorado (<http://lod.unicentro.br/Qualis/>, <http://lod.unicentro.br/SJR/> ou <http://lod.unicentro.br/SNIP/>); **(C)** exemplifica uma consulta SPARQL para consumir os recursos digitais de um determinado conjunto; e **(D)** dentre as várias formas de retorno



disponíveis, determina o formato de retorno desejado para codificação dos recursos digitais a serem recuperados.

FIGURA 3 - INTERFACE DO *ENDPOINT* HTTP://LOD.UNICENTRO.BR



Fonte: dados da pesquisa.

Considerando a Curadoria Digital, é importante destacar que o *endpoint* lod.unicentro.br tem se tornado um ponto de acesso perene aos recursos digitais (RAUTENBERG et al., 2017a; RAUTENBERG et al., 2017b). Isso privilegia os requisitos de qualidade de dados e de proveniência de dados para com as pesquisas no domínio da Ciência da Informação ao longo do tempo.

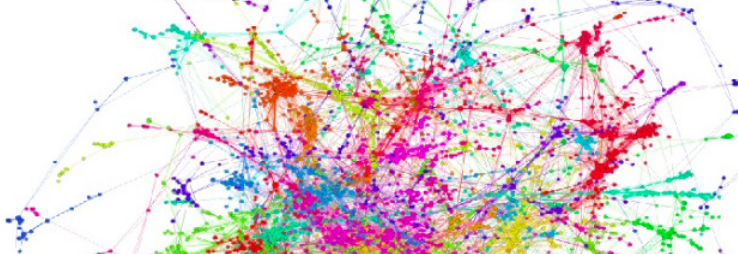


TABELA 1 - SUMARIZAÇÃO DOS RECURSOS DIGITAIS COMPARTILHADOS NO *ENDPOINT*
[HTTP://LOD.UNICENTRO.BR](http://lod.unicentro.br)

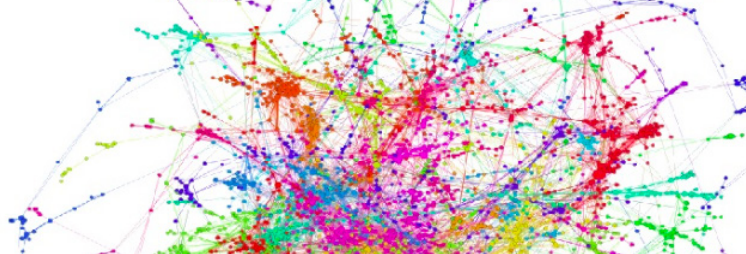
ANO REFERÊNCIA	# QUALIS	# SNIP	# SJR
2005	35.020	34.253	27.977
2006	35.020	36.342	29.570
2007	35.020	38.628	31.226
2008	54.233	41.184	32.965
2009	54.233	44.571	35.316
2010	54.233	48.484	37.997
2011	107.429	53.286	56.410
2012	107.429	56.195	59.578
2013	107.429	58.291	61.955
2014	108.622	59.888	63.629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

Fonte: Dados da pesquisa.

Destaca-se que ao se utilizar as interfaces de acesso do *endpoint*, minimiza-se os esforços de obtenção dos recursos digitais cientométricos abertamente disponibilizados, mitigando os custos da recuperação dos dados primários na atividade de Acesso, Uso e Reuso. Conforme a Tabela 1, atualmente são curados em torno de dois milhões de recursos digitais no *endpoint* lod.unicentro.br referentes aos históricos dos índices Qualis, SNIP e SJR. Tais recursos são recuperados em vários formatos de codificação, tornando-se importantes na condução e na replicação de pesquisas bibliométricas e/ou cientométricas.

4 CONSIDERAÇÕES FINAIS

Este artigo discorre a respeito de um estudo interdisciplinar e aplicado sobre Curadoria Digital e Dados Abertos Conectados. Resumidamente, relata os esforços dispendidos no desenvolvimento do *endpoint* lod.unicentro.br como fonte informacional da Web de Dados.



Ao preservar três conjuntos de recursos digitais como Dados Abertos Conectados (os históricos dos índices Qualis, SJR e SNIP), o *endpoint* lod.unicentro.br permite: **a)** a encontrabilidade dos recursos na Web de Dados; **b)** a navegabilidade entre os elementos distribuídos na web; **c)** a exploração informacional do relacionamento entre os recursos disponibilizados; e **d)** a replicação de resultados em estudos científicos ao longo do tempo.

Cientificamente, o trabalho contribui à discussão do uso interdisciplinar de Dados Abertos Conectados como arcabouço tecnológico na Ciência da Informação. Ou seja, tem-se um exemplo profícuo no que tange às novas demandas de curadoria de dados em ambientes digitais baseados na Web de Dados. Diante disso, como trabalho futuro, tem-se a contínua atualização dos históricos dos índices Qualis, SJR e SNIP no *endpoint* lod.unicentro.br.

REFERÊNCIAS

ABBOTT, D. **What is Digital Curation?** Disponível em: <<http://goo.gl/ZM6GNZ>>. Acesso em: 19 dez. 2017.

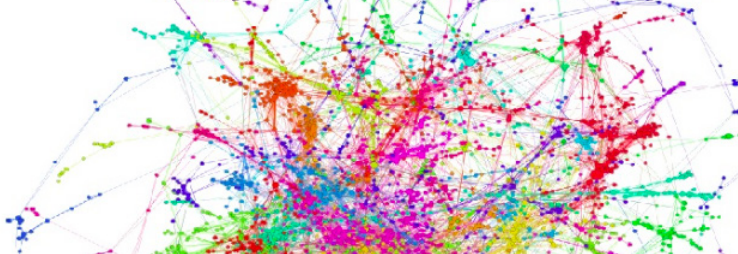
AUER, S. Introduction to lod2. In: AUER, S.; BRYL, V.; TRAMP, C (Ed.). **Linked Open Data: Creating Knowledge Out of Interlinked Data**. Springer-Verlag, 2014.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. 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/resources/curation-lifecycle-model>. Acesso em: 10 abr. 2017a.

_____. **What is digital curation?** Disponível em: <<http://www.dcc.ac.uk/digital-curation/what-digital-curation>>. Acesso em: 19 dez. 2017b.

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



LINKED DATA. **Linked Data**: design issues. Disponível em: <<http://www.w3.org/DesignIssues/LinkedData.html>>. Acesso em: 19 dez. 2017.

RAUTENBERG, S.; BURDA, A. 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.

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, 2017a.

RAUTENBERG, S.; MOTYL, S.; BURDA, A.; SILVERIO, A.; MOURA, F. M. 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.

SEMANTIC WEB. **SPARQL Endpoint**. Disponível em: <http://semanticweb.org/wiki/SPARQL_endpoint.html>. Acesso em: 19 dez. 2017.

SJR. Scimago Journal & Country Rank. Disponível em: <<http://www.scimagojr.com/journalrank.php>>. Acesso em: 20 dez. 2017.

SUCUPIRA. **Plataforma Sucupira**. Disponível em: <<http://goo.gl/Q3WA4n>>. Acesso em: 03 abr. 2017.

W3C. **Linked Data**. [2017a]. Disponível em: <<https://www.w3.org/standards/semanticweb/data>>. Acesso em: 19 dez. 2017.

_____. **Data on the Web best practices**: W3C recommendation. [2017b]. Disponível em: <<https://www.w3.org/TR/2017/REC-dwbp-20170131/>>. Acesso em: 19 dez. 2017.

_____. **Web Services Description Requirements**. [2017c]. Disponível em: <<https://www.w3.org/TR/2002/WD-ws-desc-reqs-20021028/#normDefs>>. Acesso em: 19 dez. 2017.

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