UM *WORKFLOW* AUTOMATIZADO PARA COMPARTILHAMENTO DE DADOS CIENTÍFICOS PRIMÁRIOS BASEADO EM DADOS ABERTOS CONECTADOS

AN AUTOMATED WORKFLOW FOR SHARING PRIMARY DATA BASED ON LINKED OPEN DATA

Sandro Rautenberg[1], Alessandra Cassiana Burda[2], Lucélia de Souza[3]
(1) Universidade Estadual do Centro-Oeste (UNICENTRO), R. Simeão Varela de Sá, 03 - Vila Carli, Guarapuava - PR, 85040-080, srautenberg@unicentro.br.
(2) Universidade Estadual do Centro-Oeste (UNICENTRO), R. Simeão Varela de Sá, 03 - Vila Carli, Guarapuava - PR, 85040-080, alessandra.burda@gmail.com.
(3) Universidade Estadual do Centro-Oeste (UNICENTRO), R. Simeão Varela de Sá, 03 - Vila Carli, Guarapuava - PR, 85040-080, lucelia@unicentro.br.

Resumo: Investiga a automatização dos processos para a 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* automatizado para compartilhar dados primários. Conclui-se que o *workflow* é importante na preservação de dados científicos primários, suportando tanto as pesquisas científicas quanto o reuso de recursos sob os princípios de Dados Abertos Conectados.
Palavras-chave: Dados Abertos Conectados; *Workflow*; *Workflow* para Dados Abertos Conectados; Dados Primários.

Abstract: We investigate the automation of the processes 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. As conclusion, we stand that this establishment is important for digital preservation of scientific data and can support scientific researches, considering the reuse of resources based on the Linked Open Data principles.
Keywords: Linked Open Data; Workflow; Workflow for Linked Open Data; Raw Data.

1 Introdução

A base constitutiva deste trabalho é alinhada ao que se entende por Dados Abertos Conectados (*Linked Open Data*) e as Melhores Práticas para a publicação desse tipo de recurso na *Web* de Dados. Em suma, os dados classificados como Dados Abertos Conectados são aqueles disponibilizados na *web* e regidos por licenças que advogam seu reuso por aplicações e em diversos contextos (OPEN KNOWLEDGE INTERNATIONAL, 2017; HEATH e BIZER, 2011).

Considerando as pesquisas científicas, principalmente, as que são financiadas com recursos públicos, pressupõe-se que seus dados primários devem ser compartilhados conforme os preceitos de Dados Abertos Conectados, primando pelo (re)uso de recursos em demais investigações.

Em consonância a essa visão, neste trabalho considera-se o esquema de implementação das 5 Estrelas para abertura de dados proposto por Tim Bernes-Lee. Objetiva-se o desenvolvimento de um *workflow* automatizado para publicação de dados científicos na *Web* de Dados, incrementando o grau de abertura. Neste prisma, os dados científicos devem ser publicados ao nível da 5ª Estrela (grau máximo de abertura de dados), tendo como característica principal o livre relacionamento a outros dados primários da pesquisa científica distribuídos na Internet.

Para apresentar o *workflow* proposto, além dessa seção introdutória, este artigo compreende as seguintes seções: (i) fundamentação teórica, a qual discorre sobre o conceito Dados Abertos Conectados; (ii) materiais e métodos, apontando as bases constitutiva e tecnológica do *workflow* proposto e os conjuntos de dados abertos cientométricos considerados na verficação; (iii) a apresentação do *workflow* e seus passos; (iv) verificação do *workflow*, reportando os esforços despendidos na publicação e na exploração dos índices cientométricos como Dados Abertos Conectados; e por fim (iv) considerações finais, discutindo as conclusões e trabalhos futuros.

2 Dados Abertos Conectados

Os Dados Abertos Conectados são aqueles publicados de acordo com licenças abertas, possibilitando que sejam reutilizados sem restrições, por pessoas ou aplicações e em diversos contextos. Constitutivamente, esta percepção é vinculada a dois entendimentos: (a) o que são dados abertos; e (b) o 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. Representada na Figura 1 (Apêndice A), essa classificação é denominada 5-Estrelas e é organizada como segue (5-STAR, 2017):

1ª **Estrela** - é atribuída aos dados que são 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*.

2ª **Estrela** - é conferida à publicação de dados estruturados legíveis por máquinas (*Readable Machine* - RE). Os dados são processados por *softwares* proprietários e podem ser exportados em outros formatos.

3ª **Estrela** - é concedida aos dados que são publicados em formato aberto (*Open Format* - OF). A manipulação dos dados não necessita o uso de um *software* proprietário.

4ª **Estrela** - é designada à utilização dos Identificadores Uniforme de Recursos (*Universal 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 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 dados ao fornecer uma contextualização mais ampliada.

Considerando a classificação anterior, a união de dados abertos com dados conectados é estabelecida ao se atingir a 5ª Estrela. Isso representa o ideal de publicação de dados na *web*. Ou seja, na *web*, os dados abertos podem estar conectados a outros dados, constituindo os Dados Abertos Conectados. Ressalta-se que essa união constitui a base informacional de um imenso grafo RDF, 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 (re)uso do conjunto de dados universal por diferentes pessoas e aplicações.

No contexto deste trabalho, busca-se investigar os processos para o incremento dos níveis abertura dos dados, alcançando a 5ª Estrela. Para tanto, propõem-se um *workflow* automatizado baseado no ciclo de vida *Linked Data Lifecycle* e suas tecnologias (AUER, 2014), como descrito a seguir.

3 Materiais e Métodos

A definição do *workflow* proposto é inspirada em um subconjunto das atividades do ciclo de vida *Linked Data Lifecycle* e tecnologicamente suportado pelo *Linked Data Stack* (AUER, 2014). No *workflow*, são consideradas as atividades de Extração, Armazenamento, Enriquecimento e Exploração de Dados Abertos Conectados.

No que tange a verificação do *workflow*, três bases de dados abertos do domínio da Cientometria são consideradas, sendo elas:

- **Qualis**. Segundo WebQualis (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 doze anos, a partir do Sistema WebQualis (WEBQUALIS, 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) e Rautenberg et al. (2016).

- **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 (JOURNAL METRICS, 2017). Este índice foi coletado no referido portal, em formato XLS (*eXceL Spreadsheet* - formato de planilha eletrônica da Microsoft), com o período de referência de 2005 a 2015.
- **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 (JOURNAL METRICS, 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 e fornece informações mais contextualizadas, dando 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 2015.

4 *Workflow*

Para elevar os conjuntos dados Qualis, SNIP e SJR ao nível de abertura da 5ª Estrela, um *workflow* (Figura 2 - no Apêndice B) é constituído com os seguintes passos:

- **Atividade 01 - Extração** – os arquivos em formato original são convertidos para arquivos texto. Alguns *scripts* de pré-processamento (na linguagem de programação PHP[1]) são empregados para organizar e criticar os dados.
- **Atividade 02 - Armazenamento** – os dados são armazenados em um Sistema Gerenciador de Banco de Dados Mysql[2] para serem usados por sistemas legados, por exemplo.
- **Atividade 03 – Enriquecimento** – os dados são extraídos de suas bases legadas e convertidos para arquivos no formato CSV (*Comma Separated Value*), alcançando a 3ª Estrela. Os dados também são mapeados para o formato RDF (*Resource Description Framework*) com o auxílio da ferramenta Sparqlify[3], atingindo a 4ª Estrela.
- **Atividade 04 - Armazenamento** – ao primar pelo (re)uso de dados científicos na *web*, os dados primários são compartilhados em um *endpoint* na *Web* de Dados implementado em um servidor *Open Link Virtuoso*[4] no endereço http://lod.unicentro.br/sparql (vide a Figura 3 no Apêndice C).
- **Atividade 05 – Exploração** - na *Web* de Dados, ao consultar os Dados Abertos Conectados, geralmente, objetiva a aquisição de informação contextualizada. Ao se relacionar recursos oriundos dos vários grafos RDF disponibilizados, alcança-se a 5ª Estrela.

5 Verificação

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

[1] É uma linguagem de uso geral, especialmente adequada para o desenvolvimento de aplicações *Web*. Disponível em: <http://www.php.net/>

[2] É um Sistema Gerenciador de Banco de Dados relacional *open-source* que pode ser usado em aplicações para gerir bases de dados. Disponível em: <https://www.mysql.com/>.

[3] É uma ferramenta *open-source* do Instituto *Agile Knowledge and Semantic Web* que enriquece os dados primários, convertendo os dados em triplas RDF. Disponível em: <http://aksw.org/Projects/Sparqlify.html>.

[4] Um sistema universal para acesso, integração e gerenciamento de dados baseados no modelo RDF. Disponível em: <http://virtuoso.openlinksw.com/>.

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

Para a verificação do *workflow* proposto, procedeu-se da seguinte forma. A cada índice cientométrico considerado, uma instância do *workflow* é configurada com vistas à publicação dos recursos de dados na *Web* de Dados. Nas execuções das referidas instâncias: 829.577 avaliações Qualis; 514.828 avaliações SNIP; e 485.795 avaliações SJR foram disponibilizadas. Na Tabela 1 (Apêndice D) são sumarizados os recursos de dados compartilhados, ano a ano.

Ademais, o relacionamento dos recursos disponibilizados, fomentando o alcance da 5ª Estrela, é exemplificado no Apêndice E. Na Listagem 1 do referido Apêndice, encontra-se codificada uma consulta em SPARQL que relaciona os *scores* de determinado periódico. Já a Listagem 2 exemplifica parcialmente os recursos recuperados. Ressalta-se que consultas similares ao exemplo da Listagem 1 podem ser desenvolvidas e submetidas ao *endpoint* disponibilizado. Desta forma, por exemplo, com o auxílio de APIs (*Application Programming Intefaces*), permite-se a integração dos dados abertos em outras aplicações *web*.

6 Considerações Finais

Com os estudos de caso desenvolvidos, verifica-se a adequação do *workflow* proposto para publicar dados abertos científicos na *Web* de Dados. Admite-se que este estabelecimento colabora à preservação digital de demais dados científicos primários. Inspirando-se no *workflow* desenvolvido como um modelo tecnológico, pode-se compartilhar outros recursos de dados primários, baseando-se nos preceitos de Dados Abertos Conectados.

Por isso, como trabalho futuro vislumbra-se o uso do *workflow* proposto como base para: (i) o compartilhamento de outros conjuntos de dados científicos; e (ii) a curadoria digital dos dados já disponibilizados.

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: 16 abr 2016 09:00.

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. 215p.

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

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

OPEN KNOWLEDGE INTERNATIONAL. O que são Dados Abertos? Disponível em: <http://opendatahandbook.org/guide/pt_BR/what-is-open-data/>. Acesso em: 14 jun 2017 21:00.

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.

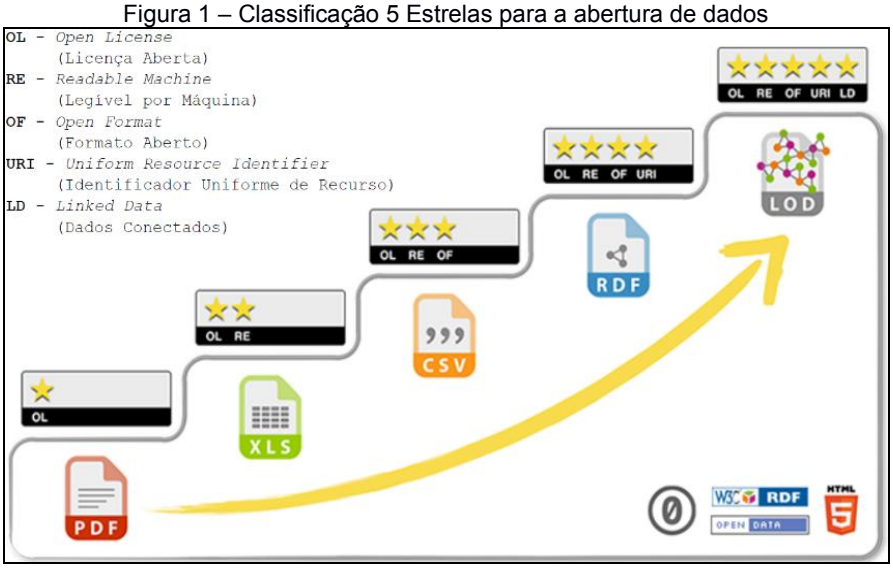
RAUTENBERG, S.; *et al*. Linked Data Workflow Project Ontology: uma Ontologia de Domínio para Publicação e Preservação de Dados Conectados. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 9, p. 1-19, 2016.

SUCUPIRA. **Plataforma Sucupira**. Disponível em: <https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf>. Acesso em: 03 abr 2017 21:00.

WEBQUALIS. **Sistema WebQualis - Portal Capes**. Disponível em: <http://qualis.capes.gov.br/webqualis/principal.seam>. Acesso em: 25 ago 2013 10:00.

Apêndice A – Classificação 5 Estrelas para a abertura de dados

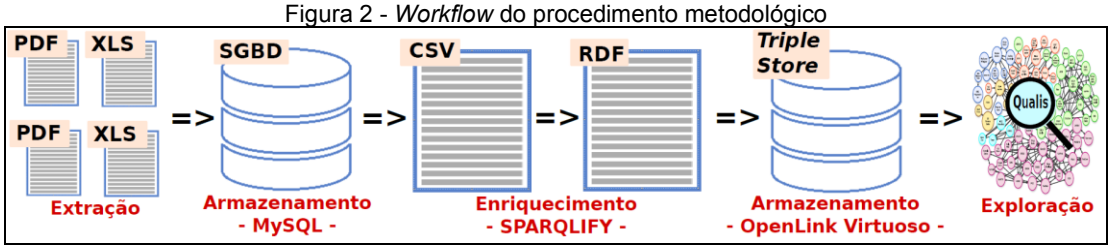
Figura 1 – Classificação 5 Estrelas para a abertura de dados



Fonte: adaptado de (5-STAR, 2017).

Apêndice B – Representação do *Workflow* automatizado

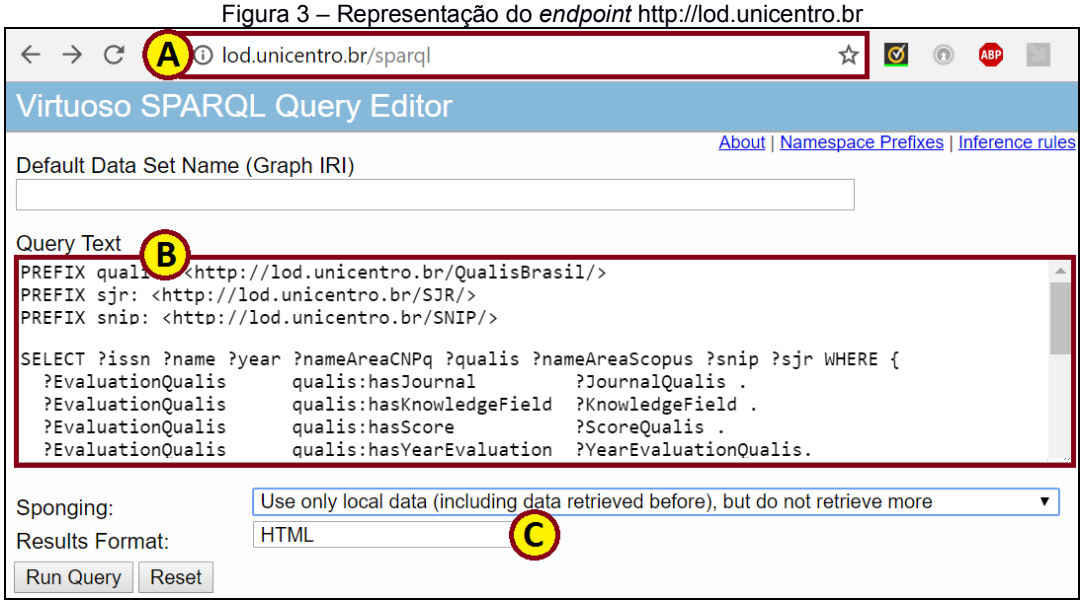
Figura 2 - *Workflow* do procedimento metodológico



Fonte: Dados da Pesquisa, 2017.

Apêndice C – Interface do *endpoint* http://lod.unicentro.br para consumo de dados

Figura 3 – Representação do *endpoint* http://lod.unicentro.br



Fonte: Dados da Pesquisa, 2017.

Apêndice C – Sumarização dos recursos compartilhados

Tabela 1 - Dados primários cientométricos compartilhados como *Linked Open Data* para pesquisas no domínio da Ciência da Informação

ANO	# AVALIAÇÕES QUALIS	# AVALIAÇÕES SNIP	# AVALIAÇÕES SJR
2005	35.020	32.932	26.881
2006	35.020	34.971	28.446
2007	35.020	37.183	30.049
2008	54.233	39.684	31.758
2009	54.233	42.984	34.074
2010	54.233	46.834	36.721
2011	107.429	51.448	54.577
2012	107.429	54.253	57.688
2013	107.429	56.360	60.019
2014	108.622	58.125	61.963
2015	44.463	60.054	63.619
2016[5]	86.446	--	--
TOTAL	829.577	514.828	485.795

Fonte: Dados da Pesquisa, 2017.

Apêndice D – Listagens da consulta e de resultado de processamento

Listagem 1 - Exemplo de consulta SPARQL que relaciona o periódico *Information Sciences* e seus índices cientométricos no ano 2015.

```
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 DISTINCT ?issn ?name ?year ?nameAreaCNPq ?qualis ?nameAreaScopus ?snip ?sjr WHERE {
      ?EvaluationQualis      qualis:hasJournal         ?JournalQualis .
10    ?EvaluationQualis      qualis:hasKnowledgeField  ?KnowledgeField .
11    ?EvaluationQualis      qualis:hasScore           ?ScoreQualis .
12    ?EvaluationQualis      qualis:hasYearEvaluation  ?YearEvaluationQualis.
13    ?JournalQualis         bibo:issn                 ?issn .
14    ?JournalQualis         foaf:name                 ?name .
15    ?KnowledgeField        dc:title                  ?nameAreaCNPq .
16    ?ScoreQualis           rdf:value                 ?qualis.
17    ?YearEvaluationQualis  rdf:value                 ?year .
18
19    ?EvaluationSJR      sjr:hasJournal         ?JournalSJR .
20    ?EvaluationSJR      sjr:hasScore           ?ScoreSJR .
21    ?EvaluationSJR      sjr:hasYearEvaluation  ?YearEvaluationSJR.
22    ?JournalSJR         bibo:issn              ?issn .
23    ?YearEvaluationSJR  rdf:value              ?year .
24    ?ScoreSJR           rdf:value              ?sjr.
25
26    ?EvaluationSNIP      snip:hasJournal         ?JournalSNIP .
27    ?EvaluationSNIP      snip:hasScore           ?ScoreSNIP .
28    ?EvaluationSNIP      snip:hasYearEvaluation  ?YearEvaluationSNIP.
29    ?EvaluationSNIP      snip:hasSubAreaScopus   ?SubAreaScopus .
30    ?SubAreaScopus       dc:title                ?nameAreaScopus .
31    ?JournalSNIP         bibo:issn               ?issn .
32    ?YearEvaluationSNIP  rdf:value               ?year .
33    ?ScoreSNIP           rdf:value               ?snip.
34    FILTER (?year = "2015" && ?issn = "0020-0255")
35  }
36
```

Fonte: Dados da Pesquisa, 2017.

Listagem 2 - Resultado parcial do processamento da consulta da Listagem 1.

```
01  "issn","name","year","nameAreaCNPq","qualis","nameAreaScopus","snip","sjr"
02  "0020-0255","Information Sciences","2015","MATEMÁTICA / PROBABILIDADE E
    ESTATÍSTICA","B1","Artificial Intelligence","2.4890","2.5130"
03  [...]
```

Fonte: Dados da Pesquisa, 2017.

[5] Quando da escrita deste artigo, as avaliações SNIP e SJR do ano 2016 não estavam disponíveis.