XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017

GT-8 – Informação e Tecnologia

PUBLICANDO DADOS NA WEB DE DADOS: UM RELATO DE EXPERIÊNCIA NA AUTOMATIZAÇÃO DOS PROCESSOS DE EXTRAÇÃO, TRANSFORMAÇÃO E CARGA DE DADOS ABERTOS PROVENIENTES DO PORTAL DADOS.GOV.BR

Sandro Rautenberg - Universidade Estadual do Centro-Oeste (UNICENTRO)

Alessandra Cassiana Burda - Universidade Estadual do Centro-Oeste (UNICENTRO)

Lucélia de Souza - Universidade Estadual do Centro-Oeste (UNICENTRO)

Josiane M. H. Dall'Agnol - Universidade Estadual do Centro-Oeste (UNICENTRO)

Gisane Michelon - Universidade Estadual do Centro-Oeste (UNICENTRO)

Tony Alexander Hild - Universidade Estadual do Centro-Oeste (UNICENTRO)

PUBLISHING DATA ON THE WEB OF DATA: AN EXPERIENCE REPORT ABOUT AUTOMATING THE EXTRACTION, TRANSFORMATION AND LOADING PROCESSES OF OPEN DATA COMING FROM THE DADOS.GOV.BR PORTAL

Modalidade da Apresentação: Comunicação Oral

Resumo: Em face ao grande volume de dados disponível na *web*, estudar as formas de preservar este volume é uma tarefa cada vez mais importante. Parte dos dados disponíveis são classificados como Dados Abertos e são disponibilizados pelos órgãos governamentais (vide o portal dados.gov.br). Esse tipo de dado tem despertado interesse para o uso em assuntos estratégicos, públicos ou privados. Neste contexto, este trabalho visa estabelecer um *workflow* para a extração de dados abertos governamentais, a transformação dos dados em dados abertos conectados e a carga dos dados transformados na *Web* de Dados, de acordo com os princípios *Linked Data* (Dados Conectados). Para tanto, utiliza-se como alicerce metodológico o ciclo de vida *Linked Data Lifecycle* e suas ferramentas. Em estudos de casos, são publicados quatro conjuntos de Dados Abertos Governamentais Conectados, os quais são originalmente provenientes do portal dados.gov.br. Com os estudos, computacionalmente, define-se um processo para a passagem dos Dados Abertos Governamentais da 3ª para 5ª Estrela, segundo a classificação de abertura de dados proposta por Tim Bernes-Lee. Como resultado desses experimentos, estabelece-se um processo automatizado como base a uma infraestrutura informacional a ser utilizada, futuramente, em um ecossistema para cidades de pequeno ou de médio porte.

Palavras-Chave: Informação Governamental. Web Semântica. Armazenamento de Dados.

Abstract: Considering the large volume of data available on the web, studying ways to preserve this volume is an important task in the context of Smart Cities. Some of the available data are classified as open data and are available from government agencies (see portal data.gov.br). This sort of data has attracted interest in strategic public or private issues. In this way, our work aims to establish an automated process for extracting open government data, transforming the data into linked open data and loading the transformed data into the Web of Data, according to the Linked Data principles. As methodological approach, this work is based on the Linked Data Lifecycle. Four case studies are developed for publishing Linked Open Government Data coming from the portal data.gov.br. With the studies, computationally, it is defined a workflow for upgrading open government data from the 3rd to the 5th Star, according to the open data classification proposed by Tim Bernes-Lee. As a result, a basilar process is established for an information infrastructure to be further applied in an Ecosystem for small or medium-sized cities.

Keywords: Government Information. Semantic Web. Data Storing.

1 INTRODUÇÃO

Considerando a atual dinâmica das Tecnologias de Comunicação e Informação, a quantidade de dados dispersos na *web* cresce exponencialmente a cada ano. Diante dessa realidade, os estudos para organizar, representar e gerenciar esse volume crescente de dados tornam-se imprescindíveis na geração de novos conhecimentos a partir da Internet (ISOTANI; BITTENCOURT, 2015).

Parte dos dados publicados na *web* são classificados como Dados Abertos. Estes são regidos por regras de livres utilização, reutilização e redistribuição por pessoas e organizações em vários contextos. Recentemente, os Dados Abertos têm despertado interesse para uso em assuntos estratégicos, tanto governamentais quanto privados. Na esfera pública, tem-se os Dados Abertos Governamentais Conectados, os quais são resultados da adoção de tecnologias da *Web* Semântica para conectar, expor e usar os dados dos sistemas governamentais (WOOD, 2011).

O Brasil é um dos países pioneiros no compartilhamento dos Dados Abertos Governamentais. Tal fato é ratificado a partir da publicação da Lei de Acesso a Informação (BRASIL, 2011), a qual inspira o portal de Dados Abertos (dados.gov.br) na disponibilização de dados sobre a prestação de contas de todas as esferas públicas. Entretanto, no referido portal somente são dispostos os Dados Abertos Governamentais que em sua origem não são conectados a outros dados na *web*.

Segundo Lebo *et al.* (2011), a publicação de Dados Abertos Governamentais Conectados, conforme os preceitos da *Web* Semântica, requer um esforço humano substancial para tornar os conjuntos de dados primários compreensíveis para os indivíduos e processáveis por computadores no ambiente distribuído da Internet. Os referidos autores pontuam que, para acelerar

o progresso na abertura de mais dados do governo, são necessárias novas abordagens para produzir a referida classe de Dados Abertos. Neste sentido, o presente trabalho tem como objetivo a discussão de uma abordagem automatizada para transformação dos Dados Abertos Governamentais brasileiros em Dados Abertos Governamentais Conectados. Como resultado desse desenvolvimento, é estabelecido um *workflow* computacional que realiza: (i) a extração de dados primários do portal dados.gov.br; (ii) a transformação dos dados primários em recursos de dados de acordo com os preceitos da *Web* Semântica; e (iii) a carga dos recursos em repositórios na *Web* de Dados.

Para discutir o *workflow* proposto, além desta seção introdutória, este artigo aborda: (i) a fundamentação teórica, estabelecendo o entendimento do conceito de Dados Abertos Governamentais Conectados; (ii) os materiais e métodos, principalmente, apontando o processo metodológico na publicação de Dados Abertos Conectados, os conjuntos de dados abertos considerados e os vocabulários usados para representar os recursos na *web*; (iii) a definição do *workflow* automatizado para a publicação dos Dados Abertos Governamentais Conectados; e (iv) as considerações finais e a apresentação dos trabalhos futuros.

2 DADOS ABERTOS GOVERNAMENTAIS CONECTADOS

Esta seção apresenta a base constitutiva dos Dados Abertos Governamentais Conectados. Abordam-se os conceitos: Dados Abertos, Dados Abertos Conectados e Dados Abertos Governamentais.

Na *web*, alguns dados publicados são classificados como Dados Abertos. Estes dados são regidos por regras claras (licenças) de livre utilização, reutilização e redistribuição por parte de pessoas e/ou organizações, nos mais variados contextos e finalidades (OPEN KNOWLEDGE INTERNATIONAL, 2017). Nesse entendimento, o termo Dados Abertos aponta para três dimensões (ISOTANI; BITTENCOURT, 2015):

- **disponibilidade e acesso** - os dados devem estar disponíveis sob custo que não seja maior que um custo razoável de reprodução, devendo estar disponível de uma forma conveniente e modificável;
- **reuso e redistribuição** - os dados devem ser fornecidos sob termos que permitam a reutilização, a redistribuição e a combinação com outros conjuntos de dados; e
- **participação universal** - os indivíduos devem ser capazes de usar, reutilizar e redistribuir os dados, sem discriminação de domínios, pessoas ou grupos.

Ao atender as dimensões anteriores, os Dados Abertos são categorizados quanto ao grau de abertura. Uma classificação conhecida é denominada "5-Estrelas" (5-STARS, 2017). Nesta classificação, quanto maior for o número de estrelas atribuído, maiores são o grau de abertura e a facilidade de interconexão a outros dados. Tal classificação é representada na Figura 1, sendo:

1ª **Estrela** - atribuída aos dados que são publicados sob licença aberta (*Open License* - OL), independente do formato em que se apresenta;

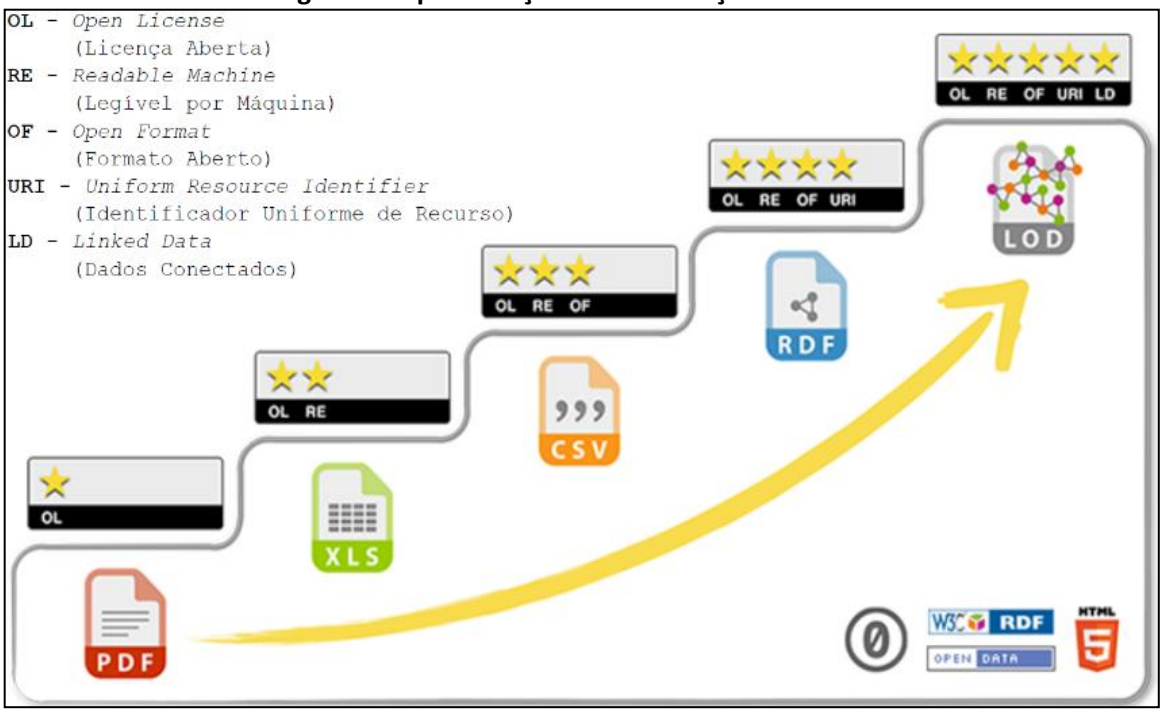
2ª **Estrela** - atribuída aos dados que além de publicados sob licença aberta são estruturados e legíveis por máquinas (*Readable Machine* - RE);

3ª **Estrela** - atribuída aos dados que são publicados em formato aberto não proprietário (*Open Format* - OF), sendo possível a manipulação dos dados sem a necessidade de uso de um *software* proprietário;

4ª **Estrela** - atribuída aos dados que possuem as classificações anteriores e que utilizam Identificadores Uniforme de Recursos (*Uniform Resource Identifier* - URI) para nomear os dados, permitindo criar ligações que façam reuso dos dados disponibilizados na *web*; e

5ª **Estrela** - atribuída aos dados que são conectados (*Linked Data* - LD) a outros dados. Permite ampliar o contexto e a descoberta de informações.

Figura 1: Representação da Classificação 5-Estrelas.



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

Ressalta-se que, considerando essa classificação 5-Estrelas, o ideal para a geração de Dados Abertos Conectados é alcançado na 5ª Estrela. Ou seja, quando os Dados Abertos estão conectados a outros dados existentes na *web*, estes compõem a classe dos Dados Abertos Conectados. Conceitualmente, tal classe de dados é regida por um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na *web*, com o intuito de subsidiar uma infraestrutura informacional global, a *Web* de Dados (BIZER; HEATH; BERNERS-LEE, 2009).

Outra classe especial de Dados Abertos são os Dados Abertos Governamentais. Esse tipo de dado é produzido, coletado ou custodiado por autoridades públicas e disponibilizados em formatos abertos (TRIBUNAL DE CONTAS DA UNIÃO, 2017). Internacionalmente, o Brasil é reconhecido como um dos pioneiros na produção e distribuição de Dados Abertos Governamentais. Conforme a Lei de Acesso à Informação, os dados sobre os indicativos e as prestações de contas de todas as esferas públicas devem estar à disposição do cidadão (BRASIL, 2011). Neste sentido, o governo federal fomenta uma plataforma de acesso aos dados governamentais públicos, o portal dados.gov.br. Neste portal, por exemplo, são publicados os dados gerados pelos Ministérios: da Educação (provas do Exame Nacional do Ensino Médio); da Agricultura, Pecuária e Abastecimento (Indicadores sobre PRONAF - Programa Nacional de Fortalecimento da Agricultura Familiar); ou do Trabalho e Previdência Social (Anuário Estatístico de Acidentes de Trabalho – AEAT). No momento da escrita deste trabalho, por exemplo, eram disponibilizados 2938 conjuntos de Dados Abertos na referida plataforma governamental (PORTAL, 2017).

Diante das classes de Dados Abertos apresentados, pode-se entender o que são os Dados Abertos Governamentais Conectados. Constitutivamente, para este trabalho, os Dados Abertos Governamentais Conectados são aqueles: (i) produzidos pelos sistemas de informações governamentais; (ii) acessados sem restrições de acordo com suas licenças de uso; e (iii) interligados a outros dados da *Web* de Dados.

3 MATERIAIS E MÉTODOS

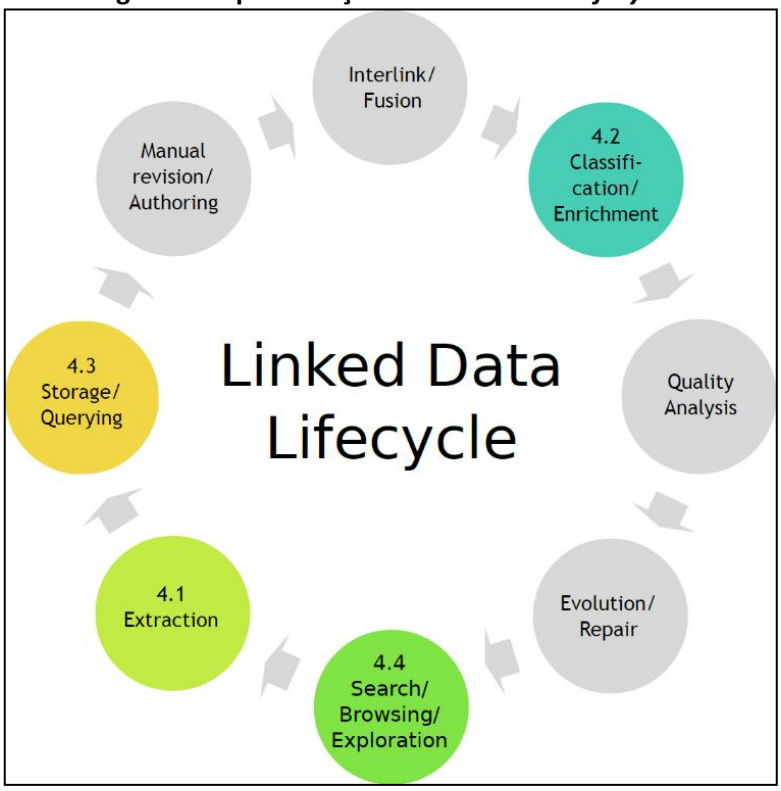
Nesta seção são abordados os insumos utilizados no estabelecimento de um *workflow* automatizado para a publicação de Dados Abertos Governamentais Conectados na *Web* de Dados. Metodologicamente, utiliza-se o *Linked Data Lifecycle*, um ciclo de vida que define as atividades para publicação de Dados Abertos. Alguns conjuntos de Dados Abertos do Ministério da

Agricultura são usados para a verificação do *workflow*. Também são apresentados os vocabulários que representam os Dados Abertos na *web* e as ferramentas utilizadas no processo de publicação.

3.1 O *Linked Data Lifecycle*

O procedimento metodológico empregado 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 aos Dados Abertos Conectados. O *Linked Data Lifecycle* consiste em oito atividades, executadas conforme os requisitos de publicação de Dados Abertos. As referidas atividades são: Extraction; Storage/Querying; Authoring; Interlinking/Fusion; Classification/Enrichment; Quality Analysis; Evolution/Repair; e Search/Browsing/Exploration. Conforme os requisitos deste trabalho, o subconjunto de atividades é destacado na Figura 2, sendo utilizadas conforme descrito na seção 4 - EXTRAÇÃO, TRANSFORMAÇÃO, CARGA E EXPLORAÇÃO DOS DADOS ABERTOS GOVERNAMENTAIS. As atividades são:

Figura 2: Representação do *Linked Data Lifecycle.*



Fonte: Adaptado de (AUER, 2014).

- **Extraction** – extrair os dados não estruturados, estruturados em diferentes formatos, ou provenientes de sistemas legados;
- **Classification/Enrichment** – utilizar as ontologias ou os vocabulários para representar os dados, suportando as atividades de recuperação de dados;
- **Storage/Querying** – utilizar as soluções computacionais para o armazenamento e a recuperação de dados no padrão RDF (*Resource Description Framework*) (PINHEIRO; FERREZ, 2014); e
- **Search/Browsing/Exploration** - empregar as soluções computacionais para consultar e/ou explorar os dados, de acordo com os objetivos do usuário.

3.2 Os Conjuntos de Dados Abertos Governamentais Utilizados

Para promover os estudos de caso de verificação do *workflow*, quatro conjuntos de dados circunscritos ao contexto do Ministério da Agricultura do Brasil são usados.

Quadro 1: Relação dos conjuntos de dados recuperados do portal dados.gov.br.

Origem dos Dados	Formato	Ano Coleta	Descrição
http://api.pgi.gov.br/api/1/serie/31.json	JSON	2017	Número de contratos firmados pelo PRONAF
http://api.pgi.gov.br/api/1/serie/32.json	JSON	2017	Valores contratados pelo PRONAF
http://api.pgi.gov.br/api/1/serie/405.json	JSON	2017	Número de contratos firmados pelo programa Mais Alimentos
http://api.pgi.gov.br/api/1/serie/407.json	JSON	2017	Valor financeiro aplicado no programa Mais Alimentos.

Fonte: Elaborado pelos autores – 2017.

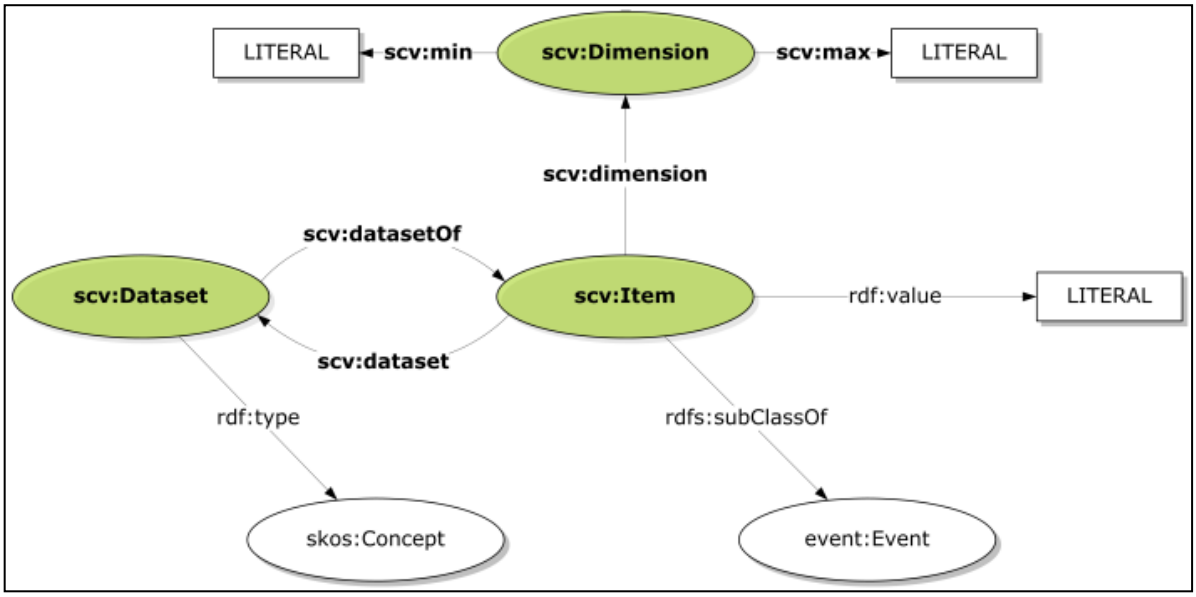
Enumerados no Quadro 1, os conjuntos de dados são provenientes do portal dados.gov.br e compreendem as séries históricas de indicadores governamentais relativos ao Programa Nacional de Fortalecimento da Agricultura Familiar (PRONAF).

3.3 Os Vocabulários Utilizados para Representar os Dados Abertos Governamentais

Para representar dados de acordo as práticas de Dados Abertos Conectados, requer-se um estudo acerca de vocabulários e ontologias existentes e comumente utilizados na *Web* de Dados. Neste sentido, o principal vocabulário a ser reutilizado é o SCOVO - *The Statistical COre VOcabulary* (SCOVO, 2017). Este se caracteriza como um vocabulário simples para representar dados estatísticos na *web*. Conforme a Figura 3, o SCOVO define três elementos principais:

- **Item -** representa uma única parte do dado;
- **Dimension -** representa parte da unidade de um dado; e
- **Dataset -** representa a unidade do dado como um todo.

Figura 3: Representação do vocabulário SCOVO.



Fonte: (SCOVO, 2017).

Outro vocabulário utilizado na *Web* de Dados é o Dublin Core (DMCI, 2017). Pontualmente, neste trabalho, utiliza-se o predicado dc:identifier para representar os identificadores de estado ou de município nos recursos de dados publicados.

3.4 As Ferramentas para Publicar os Dados Abertos Conectados

O emprego de ferramentas são parte dos esforços empreendidos frente às boas práticas para publicação de Dados Abertos Governamentais Conectados. Neste sentido, o *Linked Data Lifecycle* encontra alicerce em um conjunto de ferramentas computacionais *open-source*, o pacote *LOD2 Stack* (AUER et al., 2012). Dentre as ferramentas disponibilizadas neste pacote, destacam-se: a Sparqlify (SPARQLIFY, 2017) e a Open Link Virtuoso (VIRTUOSO, 2017).

A Sparqlify é disponibilizada na plataforma Linux e possui como característica principal 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*).

A ferramenta Open Link Virtuoso é um sistema universal para acesso, integração e gerenciamento de dados baseados no modelo RDF na *web*.

4 EXTRAÇÃO, TRANSFORMAÇÃO, CARGA E EXPLORAÇÃO DOS DADOS ABERTOS GOVERNAMENTAIS

Para publicar os quatro conjuntos de dados na *Web* de Dados, foi estabelecido um *workflow* automatizado, no qual os Dados Abertos Governamentais que estão na 3ª Estrela são elevados à 5ª Estrela, tornando-se Dados Abertos Governamentais Conectados. Os elementos deste *workflow* são descritos a seguir, de acordo com as atividades de: extração, transformação, carga e exploração de dados.

4.1 Extração de Dados

No primeiro passo do *workflow*, um conjunto de dados primários é extraído diretamente do portal dados.gov.br para um arquivo local em formato CSV. Para obter os conjuntos de dados de forma automatizada, como resultado, foi desenvolvido um *script* em linguagem PHP (acrônimo recursivo para *Hypertext Preprocessor*) parametrizável a cada conjunto de dados desejado. Conforme destacado na Listagem 1, são parametrizados:

Listagem 1: *Script* parcial em PHP para extração de dados no portal dados.gov.br.

```
01 <?php
02   $str =  file_get_contents($argv[1]);
03   $json = json_decode($str, true);
04   $fp = fopen($argv[2], 'w');
05   $array = array();
06
07   $fj = fopen("data/json/data.json", 'w');
08   fwrite($fj, json_encode($json));
09   fclose($fj);
10
11   foreach($json[$argv[3]] as $cabecalho){
12     $array = array_keys($cabecalho);
13   }
14
15   for($i = 0; $i < sizeof($array); $i++) {
16     fputs($fp, $array[$i]. ",");
17   }
18   fputs($fp, "\n");
19
20
21   foreach($json[$argv[3]] as $item) {
22     for ($i = 0; $i < sizeof($array); $i++) {
23       if(!is_null($item[$array[$i]])) {
24         fwrite($fp, $item[$array[$i]] .",");
25       }
26     }
27     fputs($fp, "\n");
28   }
29 ?>
```

Fonte: Elaborada pelos autores – 2017.

- **linha 2** - o endereço *web* do arquivo de origem em formato JSON (*JavaScript Object Notation*) que contenha o conjunto de dados primários;

- **linha 3** - uma palavra-chave que identifique o nome do registro que representa os dados a serem extraídos; e
- **linha 4** – o nome do arquivo CSV de destino a ser criado.

Salienta-se que este processo de extração mantém os dados na 3ª Estrela, preparando-os para serem convertidos para o modelo RDF (4ª Estrela).

4.2 Transformação de Dados

O próximo passo do *workflow* é a conversão do arquivo CSV para RDF. Para tanto, transforma-se os itens de dados para recursos RDF. De acordo com o ciclo de vida *Linked Data Lifecycle* essa atividade é denominada Classification/Enrichment. Para a realização desta atividade, utiliza-se a ferramenta Sparqlify usando um arquivo de configuração SML.

Listagem 2: Script SML de mapeamento do Indicador Número de Contratos PRONAF.

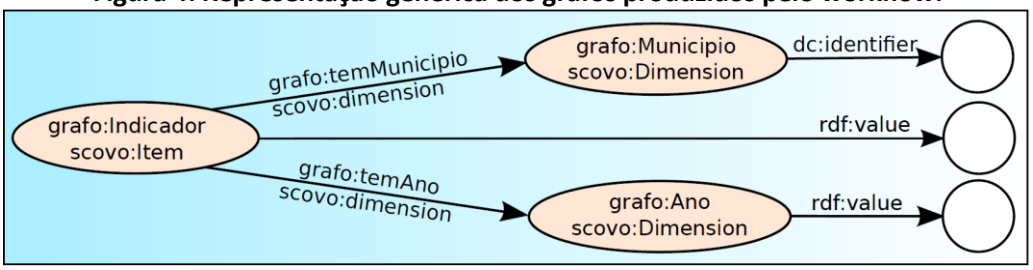
```
01  PREFIX fn:     <http://aksw.org/sparqlify/>
02  PREFIX rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
03  PREFIX dc:     <http://purl.org/dc/elements/1.1/>
04  PREFIX scovo:  <http://purl.org/NET/scovo#>
05  PREFIX scipnc: <http://lod.unicentro.br/SmartCities/PRONAF/NumeroContratos/>
06  PREFIX municipios: <http://lod.unicentro.br/SmartCities/IBGE/Municipios/>
07
08  CREATE VIEW Template DefaultView As Construct {
09    ?Item  rdf:type             rdf:Class  ;
10           rdf:type             scovo:Item ;
11           rdf:type             scipnc:Item ;
12           scovo:dimension      ?Mun ;
13           scovo:dimension      ?Ano       ;
14           rdf:value            ?valor     ;
15           scipnc:temMunicipio  ?Mun ;
16           scipnc:temAno        ?Ano       .
17
18    ?Municipio  rdf:type       rdf:Class       ;
19                rdf:type       municipios:Municipio ;
20                rdf:type       scovo:Dimension ;
21                dc:identifier  ?codMunicipio   .
22
23    ?Ano  rdf:type    rdf:Class       ;
24          rdf:type    scovo:Dimension ;
25          rdf:value   ?ano            .
26  }
27  WITH
28  ?Item = uri(concat("http://lod.unicentro.br/SmartCities/PRONAF/NumeroContratos/Item_",
              fn:urlEncode(?2),"_",fn:urlEncode(?3)))
29  ?Mun = uri(concat("http://lod.unicentro.br/SmartCities/IBGE/Municipios/Municipio_",
            fn:urlEncode(?2)))
30  ?codMunicipio = plainLiteral(?2)
31  ?Ano = uri(concat("http://lod.unicentro.br/SmartCities/Unidades/Ano_", :urlEncode(?3)))
32  ?ano = plainLiteral(?3)
33  ?valor = plainLiteral(?1)
```

Fonte: Elaborada pelos autores – 2017.

Para cada conjunto de dados, um arquivo de configuração deve ser definido. Na Listagem 2 é exemplificado o arquivo SML para mapeamento dos dados primários do número de contratos firmados no PRONAF em recursos RDF. A estrutura do arquivo SML é composta por:

- **prefixos** – no exemplo linhas 1 a 6 – define os vocabulários e ontologias utilizadas na representação dos Dados Conectados;
- **modelo RDF** – no exemplo linhas 7 a 25 – estabelece o mapeamento das colunas do arquivo CSV para com os recursos RDF; e
- **colunas CSV** – no exemplo linhas 28 a 33 – padroniza e transforma os conteúdos das colunas do arquivo CSV conforme os requisitos do RDF.

Figura 4: Representação genérica dos grafos produzidos pelo workflow.



Fonte: Elaborada pelos autores – 2017.

Considerando o exemplo da Listagem 2, a Figura 4 ilustra o modelo RDF resultante. Salienta-se que tal representação é genérica aos quatro conjuntos de dados utilizados neste trabalho. Nesta etapa do *workflow*, ao utilizar a Sparqlify, efetua-se a passagem dos dados da 3ª à 4ª Estrela.

4.3 Carga dos Dados

O passo seguinte do *workflow* consiste na carga dos dados na ferramenta Open Link Virtuoso.

Listagem 3: Script para carregamento dos grafos no servidor Open Link Virtuoso.

```
01  #!/bin/bash
02  virt_isql="isql-vt"
03  unzip_source=$1
04  virt_graphName=$2
05  virt_userName=$3
06  virt_passWord=$4
07  virt_port=1111
08
09  echo "concatenating the nt files"
10  for file in /home/ale/Documents/nt/*.nt;
11  do
12    FILESIZE=$(stat -c%s $file)
13    if [ $FILESIZE -eq 0 ];
14    then
15      echo "$file is empty" >>  errors.txt
16      $ERROR=1
17    else
18      grep "^<http" $file;
```

```
19    fi;
20  done 2>/dev/null >> /home/temp/nt/tudo.nt
21
22  echo "SPARQL CLEAR GRAPH <$2>;" | isql-vt -S "$virt_port" -U "$virt_userName"
23  -P "$virt_passWord"
24  [...]
25  # Phase 1: Unzip
26  #echo "Target: $unzip_target, Extension: $unzip_extension"
27  [...]
28  # Phase 2: Convert to n-triple
29  # FIXME Skip this step if the source file is already in n-triples format
30  [...]
31  # Phase 3: Split
32   split_source=$rapper_target
33  [...]
34  # Phase 4: Load
35  echo "creating load statement"
36  [...]
37    echo "$virt_isql $virt_port $virt_userName $virt_passWord $load_query"
38    $virt_isql "$virt_port" "$virt_userName" "$virt_passWord" "$load_query"
39  fi
```

Fonte: Elaborada pelos autores - 2017.

Com o objetivo de automatização, o armazenamento é realizado através de um *script* desenvolvido para ser executado no ambiente Linux (Listagem 3). Como parâmetros do *script*, devem ser informados no ato de sua execução:

- **linha 03** - a localização e o nome do arquivo RDF de origem;
- **linha 04** – o nome do grafo RDF a ser criado no servidor Open Link Virtuoso;
- **linha 05** – o usuário de conexão ao servidor Open Link Virtuoso; e
- **linha 06** – a senha de acesso ao servidor.

4.4 Exploração dos Dados

No consumo dos Dados Abertos Governamentais Conectados, para obter informações dos dados publicados, foram elaboradas algumas consultas na linguagem SPARQL (*Simple Protocol and RDF Query Language*), uma linguagem de consulta para extrair informações de dados baseados em triplas. Estas consultas configuram algumas informações sob o domínio da Agricultura, considerando a cidade de Guarapuava - PR.

Listagem 4: Exemplo de consulta em SPARQL para o consumo dos dados.

```
01  PREFIX dc:<http://purl.org/dc/elements/1.1/>
02  PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
03  PREFIX foaf:<http://xmlns.com/foaf/0.1/>
04  PREFIX scipnc:<http://lod.unicentro.br/SmartCities/PRONAF/NumeroContratos/>
05  PREFIX scipvc: <http://lod.unicentro.br/SmartCities/PRONAF/ValoresContratados/>
06  PREFIX municipio:<http://lod.unicentro.br/SmartCities/IBGE/Municipios/>
07
08  SELECT ?codIBGE ?nmMunicipio ?vlrAno ?qtdeContratos ?vlrContratos WHERE {
09    ?qc         rdf:type            scipnc:Item .
10    ?qc         scipnc:temAno       ?ano .
```

```
11    ?qc          scipnc:temMunicipio ?municipio .
12    ?qc          rdf:value           ?qtdeContratos .
13
14    ?qv          rdf:type            scipvc:Item .
15    ?qv          scipvc:temAno       ?ano .
16    ?qv          scipvc:temMunicipio ?municipio .
17    ?qv          rdf:value           ?vlrContratos .
18
19    ?municipio   rdf:type      municipio:Municipio .
20    ?municipio   dc:identifier ?codIBGE .
21    ?municipio   foaf:name     ?nmMunicipio .
22    ?ano         rdf:value     ?vlrAno .
23    FILTER (?nmMunicipio = "Guarapuava (PR)")
24  }
25  ORDER BY ?vlrAno
```

Fonte: Elaborada pelos autores – 2017.

Por exemplo, a Listagem 4 apresenta uma consulta que retorna o número de contratos firmados pelo PRONAF e o total de valores assumidos por esses contratos, em cada ano, na cidade de Guarapuava. O contexto ampliado é alcançado ao se relacionar informações advindas de três conjuntos de dados:

- **prefix scipnc** – linha 04 – contabiliza o número de contratos PRONAF firmados nos munícipios brasileiros, ano a ano;
- **prefix scipvc** – linha 05 – compreende os valores monetários totais dos contratos PRONAF firmados pelos munícipios, ano a ano; e
- **prefix municipios** – linha 06 – contém o nome das cidades brasileiras e o código do IBGE destas localidades.

Quadro 2: Resultado da execução da consulta da Listagem 4.

?codIBGE	?nmMunicipio	?vlrAno	?qtdeContratos	?vlrContratos
410940	Guarapuava	2000	395	2.739.411,89
410940	Guarapuava	2001	1042	2.237.019,79
410940	Guarapuava	2002	542	1.043.429,97
410940	Guarapuava	2003	280	1.092.130,67
410940	Guarapuava	2004	327	1.379.899,93
410940	Guarapuava	2005	348	1.921.974,24
410940	Guarapuava	2006	498	2.476.473,40
410940	Guarapuava	2007	440	1.551.840,25
410940	Guarapuava	2008	359	1.598.917,71
410940	Guarapuava	2009	277	2.191.134,96
410940	Guarapuava	2010	410	4.111.368,52
410940	Guarapuava	2011	380	4.564.174,76
410940	Guarapuava	2012	474	6.456.236,02
410940	Guarapuava	2013	286	3.691.361,99
410940	Guarapuava	2014	317	4.969.663,27
410940	Guarapuava	2015	279	3.903.092,30

Fonte: Elaborado pelos autores – 2017.

O

Quadro **2** apresenta o resultado da execução da consulta explicitada na Listagem 4. Neste cenário de exploração de dados, exemplifica-se como unir os dados de diversas fontes. Salienta-se que tal fato corrobora à elevação dos recursos de dados da 4ª Estrela à 5ª Estrela, ao se criar um contexto informacional mais ampliado entre os Dados Conectados disponibilizados.

5 CONSIDERAÇÕES FINAIS

Este artigo está circunscrito na interdisciplinaridade dos conceitos Dados Abertos Governamentais e Dados Abertos Conectados. Teve como propósito formalizar um *workflow* automatizado para publicação de Dados Abertos Governamentais Conectados provenientes do portal brasileiro de dados abertos. Resumidamente, os dados primários são oriundos do portal dados.gov.br e elevados da 3ª Estrela ao patamar da 5ª Estrela, segundo a classificação de Dados Abertos. O processo de publicação seguiu o ciclo de vida *Linked Data Lifecycle*. As ações que permeiam o *workflow* são aderentes às atividades de: (i) extração de dados primários; (ii) transformação de dados para o modelo RDF; (iii) carga dos recursos RDF na *Web* de Dados e (iv) exploração dos dados. Como resultado, permite-se o consumo e a exploração de Dados Abertos, alcançando a 5ª Estrela. Em resumo, com o *workflow* estabelecido, foi possível organizar, formalizar e compartilhar Dados Abertos Governamentais Conectados na *Web* de Dados, de forma automatizada.

Ademais, neste trabalho confirma-se o potencial da *Web* de Dados como uma plataforma global, em que os Dados Abertos Conectados são disponibilizados para o reuso em contextos ampliados. Observa-se como a disponibilização de recursos de dados na *web* pode ser útil no fomento de um ecossistema para cidades de pequeno ou de médio porte municiada com Dados Abertos Governamentais Conectados.

Com a experiência adquirida, como trabalhos futuros são traçados a manutenção do compartilhamento ao nível da 5ª Estrela das séries históricas trabalhadas nesta pesquisa; o início do compartilhamento de demais bases de dados circunscritas sob os demais órgãos da esfera federal; a construção de novos estudos de casos; e a remodelagem dos grafos usando o vocabulário *Model for Tabular Data and Metadata on the Web* estabelecido pelo *World Wide Web Consortium* (MODEL, 2017).

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**. 2012. Disponível em: <http://5stardata.info/en/>. Acesso em: 10 de maio 2017.

AUER, S. 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.

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. In: The Semantic Web – ISWC 2012. Cudré-Mauroux P. et al. (Ed.). **Lecture Notes in Computer Science**, v. 7650. Springer, Berlin, Heidelberg, 2012.

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

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5o, no inciso II do § 3o do art. 37 e no § 2o do art. 216 da Constituição Federal; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 18 nov. 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 15 maio 2017.

DCMI. **DCMI Metadata Terms**. Disponível em: <http://dublincore.org/documents/dcmi-terms/>. Acesso em: 27 maio 2014.

ISOTANI, S.; BITTENCOURT, I. I. **Dados Abertos Conectados: Em busca da Web do Conhecimento**. Novatec, 2015.

LEBO, T.; ERICKSON, J. S.; DING, L.; GRAVES, A.; WILLIAMS, G. T.; DIFRANZO, D.; LI, X.; MICHAELIS, J.; ZHENG, J. G.; FLORES, J.; SHANGGUAN, Z.; MCGUINNESS, D. L.; HENDLER, J. Producing and Using Linked Open Government Data in the TWC LOGD Portal. In: **Linking Government Data**. WOOD, D (Ed.). Springer-Verlag, 2011.

MODEL. **Model for Tabular Data and Metadata on the Web: W3C Recommendation 17 December 2015**. Disponível em: <https://www.w3.org/TR/tabular-data-model/>. Acesso em: 13 maio 2017.

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

PORTAL. **Portal Brasileiro de Dados Abertos**. Disponível em: <http://dados.gov.br/>. Acesso em: 16 Abril 2017.

PINHEIRO, L. V. R.; FERREZ, H. D. **Tesauro Brasileiro de Ciência da Informação**. Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), 2014.

SCOVO. **The Statistical Core Vocabulary (SCOVO) | DERI Vocabularies**. Disponível em: <http://vocab.deri.ie/scovo>. Acesso em: 16 abril 2017.

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

TRIBUNAL DE CONTAS DA UNIÃO. **5 motivos para a abertura de dados na Administração Pública**. Disponível em: <http://portal.tcu.gov.br/biblioteca-digital/cinco-motivos-para-a-abertura-de-dados-na-administracao-publica.htm>. Acesso em: 26 maio 2017.

VIRTUOSO. **OpenLink Virtuoso Home Page**. Disponível em: <https://virtuoso.openlinksw.com/>. Acesso em: 15 Maio 2017.

WOOD, D. **Linking Government Data**. Springer-Verlag, 2011.