Representação e Visualização de dados estatísticos: os desafios dos dados abertos ligados

Alan Tygel[1]

[1]Programa de Pós-Graduação em Informática – Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 68.530 – 21941-590 – Rio de Janeiro – RJ - Brasil
alantygel@ppgi.ufrj.br

Abstract. *Using a concrete study case, this work describes the work-flow for publishing statistical data represented by linked open data format. The motivation comes from data relating food production, and deaths by .. and suicide, we present the following steps in the publishing work-flow: data extraction from the original databases, data modeling using SCOVO vocabulary, triplification of the data using the Kettle software, loading triples in a triples database (Sesame Workbench), and finally, a simple script in PHP for querying and visualization of data.*

Resumo. *Através do estudo de um caso concreto, este trabalho descreve o processo de publicação de dados estatísticos representados no formato de dados aberto ligados. Utilizando como motivação dados que relacionam produção agrícola a taxas de mortalidade por câncer e suicídio, são apresentadas as seguintes etapas do fluxo de trabalho: extração dos dados das suas bases, modelamento de sua estrutura através do vocabulário SCOVO, triplificação dos dados utilizando o software Kettle, armazenamento das triplas com a plataforma Sesame Workbench, e finalmente, um script simples em PHP para consulta e visualização dos dados.*

1. Motivação

Desde de 2008, o Brasil se tornou o maior consumidor de agrotóxicos do mundo. Este fato já vem sendo tratado como um problema de saúde pública, pois os efeitos deste abuso estão se tornando cada vez mais visíveis na saúde da população do campo, no meio ambiente e nos consumidores de produtos intoxicados. [1]

O presente trabalho se desenvolve motivado pela pergunta: como utilizar os dados sobre agrotóxicos disponíveis em diversas bases de dados para gerar informações úteis no desencadeamento de ações visando frear os efeitos dos venenos agrícolas? [1]

Ao mesmo tempo, vemos hoje nos Dados Abertos Ligados (LOD, na sigla em inglês: *Link Open Data*) uma forte tendência no que tange a publicação de dados de governo no mundo inteiro.

Desta forma, este trabalho percorre o fluxo de trabalho que vai desde a obtenção

dos dados nas bases relacionais até sua consulta e visualização no formato LOD. A seções seguintes descrevem cada etapa do processo ilustrado na Figura 1. O objetivo é experimentar algumas ferramentas e vivenciar as etapas do processo de publicação.

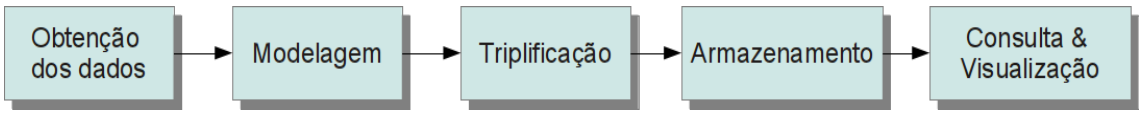


Figura 1: Fluxo de trabalho para publicação e visualização de dados em LOD

2. Extração dos Dados

Os dados utilizados neste trabalho foram fruto da pesquisa realizada em [1]. Como o objetivo desta pesquisa era dar concretude ao que foi estudado em [1], foi escolhida apenas uma parte pequeno dos dados, para que o foco se mantivesse na utilização das ferramentas.

As bases de dados utilizadas foram: Produção Agrícola Municipal (PAM/IBGE), Mortes causadas por câncer (DataSUS), Mortes causadas por suicídio (DataSUS) e População (IBGE).

Os dados de produção agrícola do IBGE são divulgados por ano, para cada municípios brasileiro, para diversas culturas permanentes e temporárias. Neste trabalho foram utilizados dados de produção de tomates e fumo, para cada município brasileiro, no ano de 2010. Os dados foram baixados do site do IBGE[1], em formato CSV.

As mortes causadas por câncer e suicídio vêm do SIM - Sistema de Informações sobre Mortalidade, ligado ao DataSus. Os dados são obtidos no formato BDF[2], acessível pelo software TabWin. A partir desta plataforma, os dados foram então transformados em formato csv. Os dados utilizados foram de mortes por câncer e suicídio, entre os anos de 1996 e 2011, nos municípios dos estados do AL, PR, RJ e RS.

Foram utilizados ainda dados sobre a população brasileira por municípios, dos ano 1980 a 2010. Os dados são de diversas pesquisas (Censo) e estimativas realizadas nos anos entre os Censos, também em formato CSV.

Como dados auxiliares, foi utilizada a base de municípios do IBGE, onde constam os nomes dos municípios e seus códigos, utilizados como identificador único. O formato desta base também é CSV.

O resumo dos dados, formatos, dimensões e atributos pode ser visto na Tabela 1.

Tabela 1: Fontes de Dados Utilizadas

Base de dados	Formato	Dimensões	Atributos
Produção Agrícola	CSV	Lavoura, Ano, Município	Unidade

1 http://www.sidra.ibge.gov.br/bda/tabela/listabl.asp?c=1612&z=p&o=35
2 http://tabnet.datasus.gov.br/cgi/sim/dados/indice.htm

Mortalidade	CSV	Doença, Ano, Município	
Municípios	CSV		Nome, código
População	CSV	Ano, Município	Unidade

3. Modelagem e Representação dos Dados

Após a definição dos dados a serem utilizados, foi feito o estudo para a representação dos dados em formato LOD. Apesar de já existir uma série de outras formas de representação mais avançadas, como o DataCube [2], optou-se pelo vocabulário SCOVO [3], por sua simplicidade de compreensão. Como trata-se de um estudo preliminar com objetivo de compreender os desafios a partir da prática, o SCOVO se mostrou a melhor ferramenta.

Este vocabulário consiste em 3 elementos principais: Item, ou seja, o dados estatístico em si; as dimensões que definem em que ponto do cubo este dado foi observado; e a base de dados a qual cada item pertence. A Figura 2 ilustra o modelo SCOVO:

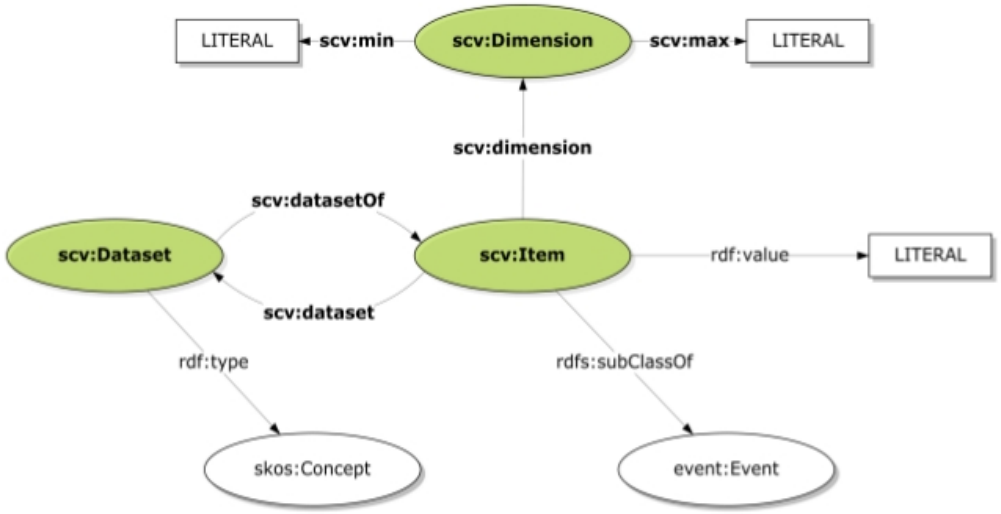


Figura 2: O Vocabulário SCOVO

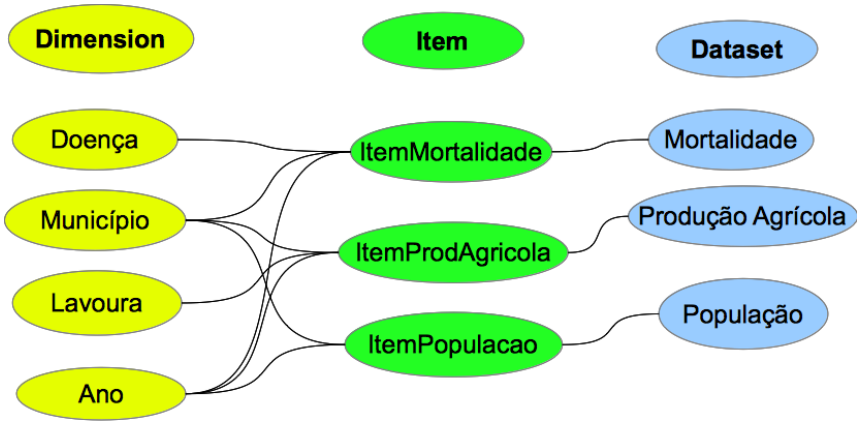


Figura 3: Problema modelado no SCOVO. Os elementos em amarelo são do tipo Dimensão, os verdes são do tipo Item, e os azuis são do tipo DataSet.

4. Triplificação

A partir da definição do modelo, a etapa seguinte foi a triplificação dos dados. Neste processo, entramos com arquivos CSV, a partir das transformações, geramos um arquivo de triplas em formato N-Triples[3].

Para esta tarefa, foi utilizado o programa Kettle, da suite Pentaho. Esta ferramenta permite que o processo de triplificação seja automatizado. Em geral, os passos são os seguintes:

1. Leitura do arquivo e definição dos campos que serão utilizados;
2. Para cada linha, geram-se as URIs do item, também aquelas relacionadas às dimensões.
3. A partir delas, geramos uma frase na linguagem Turtle. Com ela, enunciamos uma vez a URI do sujeito, e depois escreve-se somente as URIs dos predicados e objetos. Para cada item, temos um tripla com com o valor daquela observação, o tipo (item), uma tripla relacionando o item ao Dataset, uma tripla para cada dimensão, além de uma tripla descrevendo atributos do item, como por exemplo uma unidade de medida.

Figura 4: Kettle: Leitura do arquivo original, seleção dos campos, inserção do prefixo, construção de cada URI, construção das frases em Turtle, seleção apenas das frases, e escrita do arquivo.

3 http://www.w3.org/2001/sw/RDFCore/ntriples/

5. Armazenamento

Com as triplas geradas, podemos então inseri-las em um banco de triplas. O sistema escolhido foi o Sesame Workbench, escrito em Java. Sua instalação é simples, bastando um servidor TomCat rodando.

O Sesame permite armazenar as triplas na memória (arquivos texto) ou em SGBDs relacionais, como o MySQL. Para este experimento foi escolhida a opção de arquivos texto, entretanto vê-se que a performance piora à medida que a quantidade de dados aumenta.


Explore (<http://dados.contraosagrotoxicos.org/Resource/Producao/Tom

Subject	Predicate	Object	Context
<http://dados.contraosagrotoxicos.org/Resource/Producao/Tomate-110001-2011>	<http://dados.contraosagrotoxicos.org/property/Lavoura>	<http://dados.contraosagrotoxicos.org/Resource/Lavoura/Tomate>	<http://ww
<http://dados.contraosagrotoxicos.org/Resource/Producao/Tomate-110001-2011>	<http://dados.contraosagrotoxicos.org/property/Municipio>	<http://dados.contraosagrotoxicos.org/Resource/Municipio/110001>	<http://ww
<http://dados.contraosagrotoxicos.org/Resource/Producao/Tomate-110001-2011>	<http://dados.contraosagrotoxicos.org/property/Ano>	<http://dados.contraosagrotoxicos.org/Resource/Ano/2011>	<http://ww
<http://dados.contraosagrotoxicos.org/Resource/Producao/Tomate-110001-2011>	<scovo:Dataset>	<http://dados.contraosagrotoxicos.org/Resource/DataSet/ProducaoAgricola>	<http://ww
<http://dados.contraosagrotoxicos.org/Resource/Producao/Tomate-110001-2011>	<http://dados.contraosagrotoxicos.org/property/Unidade>	"ton"	<http://ww
<http://dados.contraosagrotoxicos.org/Resource/Producao/Tomate-110001-2011>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://dados.contraosagrotoxicos.org/Resource/ProducaoAgricola>	<http://ww
<http://dados.contraosagrotoxicos.org/Resource/Producao/Tomate-110001-2011>	<http://dados.contraosagrotoxicos.org/property/Valor>	148	<http://ww

Figura 5: Sesame: Interface de navegação pelas triplas

O Sesame permite que as triplas sejam inseridas pelo arquivos texto gerados pelo Kettle. Com as triplas dentro do Sesame, é possível acessar seu *end-point* e fazer consultas SPARQL, tanto por dentro da interface quanto através de pedidos REST. Esse foi o procedimento adotado na etapa seguinte.

6. Consulta e Visualização

A linguagem SPARQL não é uma linguagem que possamos exigir que um usuário de um sistema domine. Portanto, devemos construir consultas pré-moldadas para que a nossa base de triplas possa ser acessada.

Neste trabalho, foram construídas três consultas: A primeira lista todas as dimensões e suas instâncias criadas no modelo. Pode-se ver (Figura 6), portanto, cada tipo de Lavoura, cada tipo de Doença, e todos os municípios inseridos como dimensões.

A segunda consulta mostra uma tabela com todos os municípios, o número de ocorrências de mortes por câncer e suicídio no ano de 2010, e a colheita de fumo e tomate no ano de 2011.

A terceira consulta lista a mesma tabela, mas permite ordenar por cada tipo de lavoura ou por cada tipo de doença (Figura 7).

Com isso, qualquer usuário pode facilmente acessar a base de dados e extrair algumas informações. No entanto, esse acesso fica limitado às consultas pré-definidas.

O programa foi escrito em PHP, e as consultas foram feitas da forma mais genérica possível, de modo que não precise ser alterada em caso de entrada de mais dimensões ou dados. O código encontra-se em arquivo anexo.

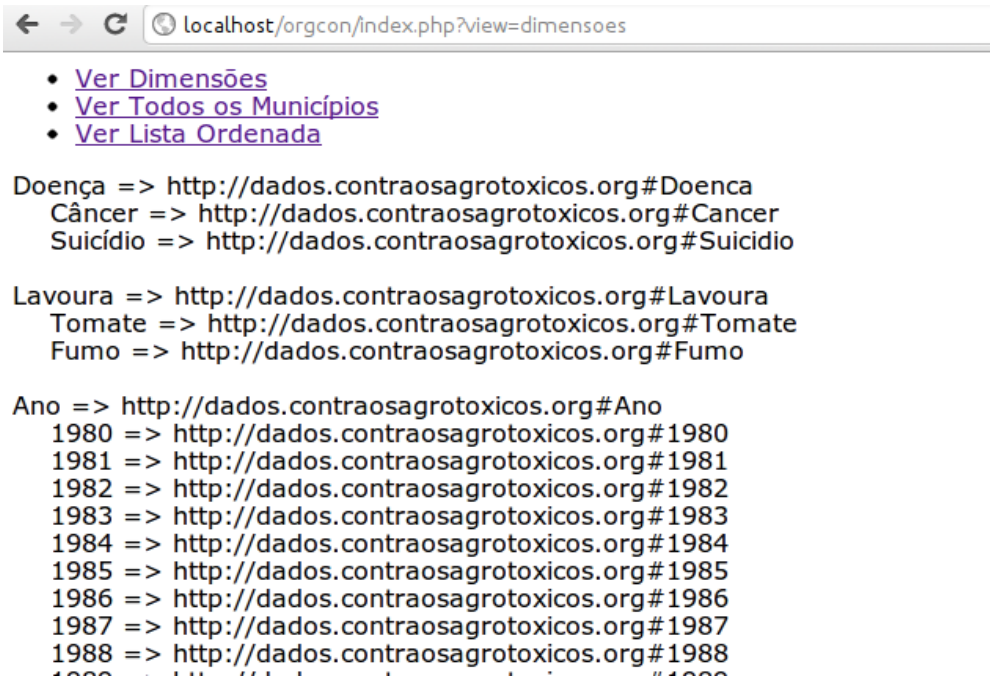
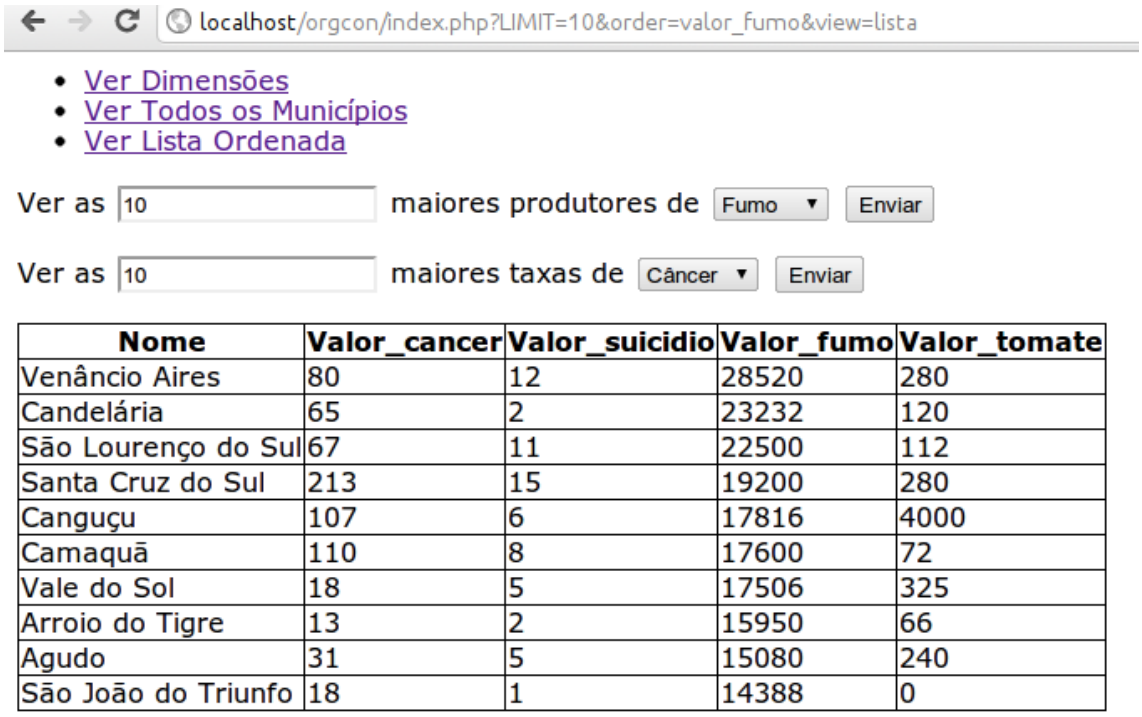


Figura 6: Listando as dimensões e suas instâncias. Os municípios não aparecem na imagem.



Nome	Valor_cancer	Valor_suicidio	Valor_fumo	Valor_tomate
Venâncio Aires	80	12	28520	280
Candelária	65	2	23232	120
São Lourenço do Sul	67	11	22500	112
Santa Cruz do Sul	213	15	19200	280
Canguçu	107	6	17816	4000
Camaquã	110	8	17600	72
Vale do Sol	18	5	17506	325
Arroio do Tigre	13	2	15950	66
Agudo	31	5	15080	240
São João do Triunfo	18	1	14388	0

Figura 7: Listando os maiores municípios produtores de fumo.

7. Conclusões

O objetivo deste trabalho não foi fazer uma publicação completa de dados aberto ligados. Foi dada maior importância no estudo da ferramenta e na análise da viabilidade de trabalho com dados em grande escala. Neste sentido, é necessário listar os problemas encontrados no percurso, dando a dimensão de quais etapas ainda seriam necessárias para uma publicação completa:

1 – **Capacidade do Sesame.** Não foi possível inserir o dados de população na plataforma Sesame, pois elas superavam o limite de memória estabelecido no programa. Pelo mesmo motivo, só foi utilizadas a produção agrícola de um ano. Não foi possível alterar este limite de memória.

2 – **Ligação com outras bases.** Os dados aqui trabalhados não foram ligados com outras bases. Exemplos de bases com as quais poderíamos compartilhar os vocabulários são: AGROVOC[4], um vocabulário gerido pela *Food and Agriculture Organization* (FAO, ligada à ONU), poderia ligar as lavoura e a produção de alimentos; Vocabulários para dados geográficos, em que poderiam ter sido colocados os municípios.

3 – **Completude dos dados.** Foram utilizados apenas alguns estados, quando havia dados disponíveis para mais estados.

4 – **Visualização.** Não foi feito nenhuma visualização gráfica além de tabelas. Utilizando alguma biblioteca de gráficos em PHP, poderia ser feita uma visualização da evolução da mortalidade por câncer e suicídio nos municípios ao longo dos anos, junto com a evolução da produção agrícola.

Desta forma, podemos concluir que o trabalho foi útil para experimentar o processo de publicação de dados e perceber os limites que ainda existem neste campo. Certamente, um trabalho de pesquisa mais aprofundado pode resolver os problemas de performance. A modelagem utilizando o vocabulário SCOVO, apesar de simples, mostrou-se capaz de resolver o problema colocado.

Bibliografia

[1] Tygel, Alan. *Dados sobre Agrotóxicos: Informação para Ação.* Relatório Final da Disciplina de Fundamentos de Modelagem, professoras Maria Luiza Machado Campos e Jonice Oliveira. PPGI/UFRJ, 1. trimestre de 2012.

[2] Cyganiak, R., Field, S., Gregory, A., Halb, W. & Tennison, J. Semantic Statistics: Bringing Together SDMX and SCOVO. *LDOW* **628**, (2010).

[3] Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L. & Ayers, D. SCOVO: Using Statistics on the Web of Data. *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications* 708–722 (2009).

4 http://aims.fao.org/standards/agrovoc/linked-open-data