Aplicando *Linked Data* na publicação de dados do ENEM

Samuel Pierri Cabral[1], Nitay Batista Beduschi[1], Airton Zancanaro[2], José Leomar Todesco[12], Fernando A. O. Gauthier[12]

[1]Departamento de Informática e Estatística– Universidade Federal de Santa Catarina (UFSC) – Florianópolis, SC – Brasil

[2]Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento – Universidade Federal de Santa Catarina (UFSC) – Florianópolis, SC – Brasil

{scabral, nitay}@inf.ufsc.br, {airtonz, tite, gauthier}@egc.ufsc.br

Abstract. *The present article describes the experiment performed with the National Examination of Secondary Education (ENEM) data set, in the year of 2008, which were published in the principles of linked data. To this end, the data set were first treated in a database, performed consolidation and represented by an ontology. Then, with the use of tools, the data set were converted to a RDF format, linked to the data of DBPedia and published in a triple store. Lastly, a Web application was built to allow visualization of the data with the aid of SPARQL consultations. The experiment allowed us to establish a workflow for publication of linked geographical data, allowing of analysis and discovery of new knowledge.*

Resumo. *Este artigo descreve o experimento realizado com os dados do Exame Nacional do Ensino Médio (ENEM), do ano de 2008, que foram publicados nos princípios do Linked Data. Para tal, os dados foram primeiramente tratados em um banco de dados relacional, realizadas consolidações; e os dados, representados por uma ontologia. Em seguida, com o uso de ferramentas, foram convertidos para o formato RDF, ligados aos dados da DBPedia e publicados em um servidor de triplas. Por fim, uma aplicação Web foi construída para a visualização dos dados com auxílio de consultas SPARQL. O experimento permitiu estabelecer um fluxo para publicação de dados geográficos, possibilitando a descoberta de novos conhecimentos.*

1 Introdução

O Ministério da Educação (MEC) tem aplicado anualmente o ENEM, cujos resultados, para o governo, são tidos como um instrumento de avaliação e promoção de melhorias na educação e, para o secundarista, como a possibilidade de acesso às universidades públicas. Os dados obtidos com o exame ficam disponíveis para que a sociedade converta-os de forma tal que sejam entendidos também pelas máquinas e possam ser ligados a outros conjuntos de dados, permitindo complemento das informações.

Na tentativa de avaliar o desempenho do secundarista e a qualidade do ensino médio, o MEC criou em 1998, o ENEM (CASAGRANDE, 2009). Posteriormente, o exame passou por reestruturações e foi utilizado, também, como forma de acesso de novos estudantes a universidades públicas brasileiras através do SiSU (Sistema de

Seleção Unificada). Em 2011, segundo Carneiro (2011), foi considerado o maior exame existente na América Latina, contando com mais de 4,5 milhões de inscritos.

No ENEM, cinco competências são avaliadas, tanto nas provas objetivas como nas subjetivas (INEP, 2008). São elas: 1) dominar linguagem; 2) compreender fenômenos; 3) enfrentar situações-problema; 4) construir argumentação; e 5) elaborar propostas. Isso permite, por exemplo, que cada instituição de ensino possa criar seus próprios critérios para o acesso de novos estudantes.

Todas as informações relacionadas ao exame estão disponíveis para ser acessadas livremente através do *site* do INEP[1], que disponibiliza também, o "Manual do usuário dos microdados do ENEM". Isso permite que, através desses dados, diferentes pessoas possam acessá-los, distribuí-los, reusá-los, analisá-los ou publicá-los usando os princípios do *Linked Data*.

Este surgiu em 2006 (BERNERS-LEE, 2006) como uma alternativa para tentar resolver o problema da grande quantidade de dados disponíveis na *Web*, isto é, textos ou arquivos dos mais variados formatos, dos quais muitos só podem ser interpretados por seres humanos (BERNERS-LEE, 2010). Isso impede que aplicações (máquinas) consigam extrair informações reais contidos nesses documentos.

Assim como os *hiperlinks,* que permitem conectar documentos em um espaço único de informação global, Heath e Bizer (2011) afirmam que o *Linked Data* possibilita a ligação entre diferentes fontes de informação, formando a *Web* de dados. Isso torna possível que as aplicações genéricas operem sobre um conjunto de dados mais completo.

Pesquisas, como a de Hull (1997), discutem a integração das diferentes bases de dados, tendo o propósito de agregar mais conteúdo ao que está sendo pesquisado na *Web*. No que se refere a dados estatísticos, Zapilko e Mathiak (2011) e Kämpgen, O´Rain e Harth (2012) apresentam o formato de cubo OLAP como método para aumentar a *performance* na publicação de dados na *Web*. Já Pirrotta (2010) descreve o processo e as lições aprendidas na publicação dos dados das universidades utilizando os princípios do *Linked data*.

Para que a interligação dos dados seja possível, Bizer, Heath e Berners-Lee (2009) descrevem um conjunto de regras, conhecidas como os quatro princípios do *Linked Data*, e estabelecem um padrão para a publicação dos dados na *Web*. São elas: 1) utilizar *Uniform Resource Identifiers* (URI) para nomear as coisas; 2) usar HTTP URIs para que as pessoas possam procurar por esses nomes; 3) fornecer informações úteis utilizando os padrões *Resource Description Framework* (RDF) e *SPARQL*, quando alguém procurar por uma URI; e 4) incluir *links* para outras URIs a fim de que elas possam descobrir mais informações.

Dessa forma, as aplicações interpretam os conteúdos disponibilizados na *Web* de dados através de um modelo genérico, denominado de RDF, que se liga a outros dados no mundo na forma de triplas: sujeito, predicado e objeto (LASSILA; SWICK, 1998). A proposta do RDF é, segundo Souza e Alvarenga (2004), criar uma maneira com a qual cada página *Web*, cada recurso possam gerar sua própria metainformação, ou seja, informação sobre informação, e torná-la disponível para quem precisar.

[1] INEP: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

Os padrões RDF, associado às ontologias e aos *namespaces* compartilhados, possibilitam combinar os dados, como no caso do ENEM, com outras fontes de dados, buscando, assim, mais informações referentes, p. ex., ao município onde foi realizada a prova: sua latitude e longitude, informações sobre seu IDH e sua população etc.

Em síntese, este trabalho tem por objetivos identificar os dados do ENEM disponíveis para serem acessados de forma aberta, tratá-los conforme a necessidade, representá-los por meio de ontologias, convertê-los no formato RDF e construir uma aplicação *Web* a fim de visualizá-los de forma amigável. Para isso, na seção 2, será apresentada a definição do escopo do trabalho; na seção 3, serão abordados o tratamento dos dados e a conversão para o formato RDF; na seção 4, são demonstrados os resultados; e, por último, na seção 5, apresentadas as considerações finais.

2 Definição do escopo do trabalho

Para a elaboração deste experimento, inicialmente procurou-se determinar quais informações relativas ao ENEM estavam disponíveis para utilização no *site* do INEP. Essa consulta foi realizada no mês de outubro de 2011, e identificou-se que, além de microdados de outros anos, o mais recente era 2008, utilizado no trabalho.

Com o auxílio do manual do usuário, percebeu-se que os dados estão estruturados em variáveis de controle do inscrito e da escola, da cidade da prova, da prova objetiva e de redação, e do questionário socioeconômico.

Dentre tantas variáveis, optou-se, para o tratamento e a publicação, agrupá-las por município em que o inscrito realizou a prova, visto que grande parte das informações nesse campo estava preenchida adequadamente. Com esse agrupamento foi possível criar indicadores das médias obtidas na prova objetiva e na redação, do sexo, da idade, da rede de ensino e da localização da escola (área urbana ou rural). Além disso, a criação de uma ontologia e o reuso de outra se fizeram necessários, permitindo que os dados do ENEM pudessem ser interligados a outros através do padrão RDF.

3 Tratamento e publicação dos dados

Os dados do ENEM de 2008 estão no formato texto e com as informações separadas por ponto e vírgula. Com o auxílio das ferramentas UltraEdit, para a conversão do arquivo no formato CSV (*Comma Separated Values*), e do MySQL Workbanch, foi possível criar as tabelas e importar os dados para o banco de dados MySQL.

De posse das informações em uma tabela no banco de dados, denominada de "CSV_ENEM", uma nova foi criada ("EN_INDICADORES_ENEM"), com os cálculos dos indicadores agrupados por município, juntamente com os outros campos descritos anteriormente, utilizados para a geração do RDF. O Quadro 1 exemplifica o tratamento realizado em um campo da tabela.

Quadro 1 – Exemplo de um campo da tabela "EN_INDICADORES_ENEM"

Campos da tabela	Descrição dos campos	Tratamento realizado
COD_MUNICIPIO	Código do IBGE do município onde foi realizada a prova	No campo ID_CIDADE_PROVA da tabela CSV_ENEM, foi utilizada parte do código para compor o código do IBGE; os municípios sem código ou com código sem equivalência aos do IBGE foram alterados para '0'.

Para atender ao escopo deste experimento, foi criada uma ontologia, denominada de "enem[2]", com o objetivo de fazer a interligação dos dados do exame a outros existentes. Essa ontologia foi criada a partir da importação da ontologia "geopoliticabr"[3], disponível no projeto Lodkem[4], que possui as propriedades referentes aos municípios brasileiros, como o código do IBGE, o nome, o *sameAs* para a DBpedia e os pontos cardeais. A ontologia "geopoliticabr" pode ser visualizada na Figura 1.

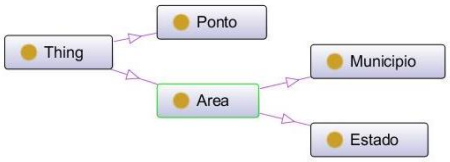

Figura 1 - Ontologia "geopoliticabr"

Com a ontologia disponível na *Web,* para ser acessada, e os dados tratados em uma tabela de banco de dados, foi possível construir um mapeamento entre eles, de forma tal, que o *software* D2RQ tivesse condições de gerar o RDF. Isso permitiu que, importando o RDF para um servidor de triplas (OpenLink Virtuoso), os dados se tornassem disponíveis, ligados a outros e a DBPedia, possibilitando a consulta através da linguagem SPARQL e a construção de uma aplicação.

4 Demonstração dos resultados

Para facilitar a visualização dos resultados, uma aplicação *Web* foi criada e disponibilizada no *site* http://www.lodkem.ufsc.br:8080/enem, com a finalidade de construir gráficos dinâmicos, utilizando as consultas SPARQL. A consulta de um município é exemplificada no Quadro 2.

Quadro 2 - Consulta SPARQL do município de Florianópolis

```
PREFIX geo: <http://lodkem.ufsc.br/onto/geopoliticabr#>
PREFIX enem: <http://lodkem.ufsc.br/onto/enem#>
SELECT DISTINCT ?nome ?codibge ?lat ?lon ?area ?totalaluno
WHERE {
  ?x geo:temNomeMun ?nome .
  ?x geo:temCodIbgeMun ?codibge .
  ?x geo:temPontoCentralMun ?p .
  ?x geo:temAreaMun ?area .
  ?p geo:lat ?lat .
  ?p geo:lon ?lon .
  ?x enem:temTotalAluno
  ?totalaluno .
  filter (?codibge != 0 && regex(?nome,'florianopolis', 'i'))
} ORDER BY (?codibge)
```

Através dessa consulta, os dados do município, como o nome, o código do IBGE e a latitude e longitude, são buscados através da ontologia "geopoliticabr". Já o total de alunos são oriundos da ontologia "enem". Além disso, o filtro "regex" permite trazer todos os municípios com o nome informado, e a *string* 'i' significa *case insensitive*. O campo e os dados geográficos da cidade pesquisada podem ser observados na Figura 2.

[2] Disponível em: http://lodkem.ufsc.br/onto/enem.owl
[3] Disponível em http://lodkem.ufsc.br/onto/geopoliticabr.owl
[4] http://lodkem.egc.ufsc.br/

Para gerar o gráfico com as médias das cinco competências avaliadas na prova objetiva e a média geral, é executada a seguinte consulta (Quadro 3):

Quadro 3 - Consulta SPARQL com a média das competências

```
PREFIX enem: <http://lodkem.ufsc.br/onto/enem#>
PREFIX geo: <http://lodkem.ufsc.br/onto/geopoliticabr#>
SELECT DISTINCT ?nome ?codibge ?compobj1 ?compobj2 ?compobj3 ?compobj4
?compobj5 ?mediageralobj
WHERE {
   ?x geo:temNomeMun ?nome .
   ?x geo:temCodIbgeMun ?codibge .
   ?x enem:temMediaCompetenciaObj1 ?compobj1 .
   ?x enem:temMediaCompetenciaObj2 ?compobj2 .
   ?x enem:temMediaCompetenciaObj3 ?compobj3 .
   ?x enem:temMediaCompetenciaObj4 ?compobj4 .
   ?x enem:temMediaCompetenciaObj5 ?compobj5 .
   ?x enem:temMediaGeralObj ?mediageralobj .
   filter (?codibge = 4205407 ) }
```

Essa consulta tem como resultado o gráfico apresentado na Figura 3. Vale ressaltar que o foco deste trabalho não é analisar os resultados cognitivos dos estudantes, e sim, a disponibilização e visualização das informações.

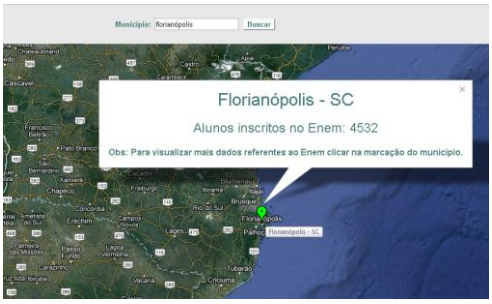

Figura 2 - Mapa com o resultado da consulta por município

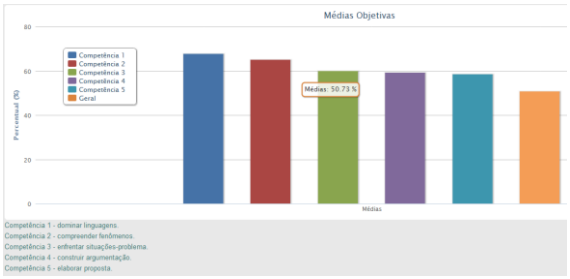

Figura 3 - Média das cinco competências da prova objetiva e média geral

Desta forma, percebe-se que, com a ligação dos dados a outras fontes, é possível descobrir facilmente mais informações, ampliando o conhecimento sobre determinado assunto.

5 Considerações finais

Este trabalho buscou identificar os dados do ENEM que estão disponíveis de forma aberta, fazer o seu tratamento, construir e reusar ontologias, convertê-los no formato RDF e desenvolver uma aplicação *Web* que lhes permitisse a visualização de uma forma amigável.

Quanto às lições aprendidas, concernentes à publicação de dados abertos, viu-se que, de um modo geral, o governo vem disponibilizando uma quantidade significativa de dados. Entretanto, a falta de estruturação e dados incompletos dificultaram a sua manipulação de forma adequada.

Para trabalhos futuros, será realizada a extensão da ontologia "geopoliticabr", incorporando novos conceitos e a inclusão de dados do ENEM dos outros anos. Além disso, para aumentar a escala de publicação e análise através de séries históricas dos dados, o vocabulário *Data Cube* RDF está sendo automatizado.

Dessa forma, a contribuição deste experimento está na utilização de métodos e ferramentas para a publicação de dados utilizando os princípios do *linked data.* Acredita-se que, com infraestrutura de acesso aos dados padronizados e em escala global, será possível torná-los disponíveis tanto para o uso humano quanto para as máquinas. Isso facilitará para que as aplicações construídas tenham condições de reutilizar os dados facilmente, formando uma base fundamentada para novos conhecimentos.

Referências

BERNERS-LEE, Tim. **Linked Data: Design Issues.** 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 05 jul. 2012.

_____. Long Live the Web: A Call for Continued Open Standards and Neutrality. **Scientific American**, Dezembro 2010. Disponível em: <http://www.scientificamerican.com/article.cfm?id=long-live-the-web&page=6>. Acesso em: 29 Abr. 2012.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data - The Story So Far**. Int. Journal On Semantic Web And Inf. Systems (ijswis),** p. 1-22. mar. 2009.

CARNEIRO, V. L. Políticas Públicas Educacionais e Gestão do Ensino Médio no Brasil: o Exame Nacional de Ensino Médio - ENEM e suas implicações para o trabalho docente. In: **XXV Simpósio Brasileiro e II Congresso Ibero-Americano de Política e Administração da Educação**, 2011, São Paulo.

CASAGRANDE, A. L. Avaliação: a redefinição do papel do ENEM. In **XI Seminário Estadual da ANPAE-SP**, 2009.

INEP. **Matriz De Referência Para O Enem**. Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira. Brasília, p. 26. 2008.

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

HULL, R. "Managing Semantic Heterogeneity in Databases: A Theoretical Perspective," in **ACMSymposium on Principles of Databases**, 1997, pp. 51–61.

KÄMPGEN, B.; O'RAIN, S.; HARTH, A. **Interacting with Statistical Linked Data via OLAP Operations. in** Inter. Workshop on Interacting with Linked Data. 2012.

LASSILA, O.; SWICK, R. R. **Resource Description Framework (RDF) Model and Syntax Specification**. Disponível em: <http://www.w3.org/1998/10/WD-rdf-syntax-19981008>. Acesso em: 28 ago. 2012.

PIRROTTA, G. Linking Italian University statistics. **ACM International Conference Proceeding Series**, 2010. Graz.

SOUZA, R. R. ; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, Brasília, v. 33, n. 1, p. 132-141, 2004.

ZAPILKO, B.; MATHIAK, B. Performing Statistical Methods on Linked DataProc. Int'l Conf. on Dublin Core and Metadata Applications 2011. Anais...2011.