# Um *Linked Data Mashup* de Dados de Execuções Financeiras e Indicadores Educacionais no Ensino Básico

**Caio Viktor da Silva Avila[1], Tulio Vidal Rolim[1], Matheus Mayron Lima da Cruz[1], Amanda Drielly Pires Venceslau[1], José Wellington Franco da Silva[1], Vânia Maria Ponte Vidal[1]**

[1]Departamento de Computação – Universidade Federal do Ceará (UFC)
Fortaleza, CE – Brasil

{arlaass,tulio.xcrtf,matheusmayron,jwellingtonfranco,driellyads
vaniap.vidal}@gmail.com

***Abstract.*** *The purpose of this work is to use Linked Data and Semantic Web techniques to construct a mashup with financial execution data from PDDE and educational indicators. At the end of the process an RDF dataset was generated containing an integrated view of the sources.*

***Resumo.*** *O objetivo deste trabalho é utilizar técnicas de Dados Ligados e Web Semântica para a construção de um mashup com dados de execução financeira do PDDE e de indicadores educacionais. Ao final do processo foi gerado um dataset RDF contendo uma visão integrada das fontes.*

## 1. Introdução

A qualidade da educação é influenciada por diversos fatores, onde dentre esses, os investimentos destacam-se como um meio de contribuir na promoção de um padrão de qualidade mínimo ao ensino, bem como reduzir a discrepância em seu acesso. O investimento de recursos financeiros estimula a boa qualidade na relação ensino-aprendizagem, trazendo um impacto positivo nas escolas, assim, o investimento para melhoria na estrutura das escolas apresenta-se como um meio de fornecer uma melhor forma de aprendizagem aos estudantes, através de práticas e atividades interativas [Júnior et al. 2017]

O Programa Dinheiro Direto na Escola (PDDE) [FNDE 2017] é um dos programas de promoção ao investimento financeiro que busca contribuir na manutenção e aprimoramento da infraestrutura física e pedagógica das escolas. Os dados sobre Execução Financeira do PDDE são disponibilizados através do Plano de Dados Abertos do FNDE [FNDE 2018]. De forma semelhante, os dados sobre taxas de rendimento de estudantes e as médias de alunos por turma também são disponibilizados no Portal do INEP [INEP 2018]. Nessa conjuntura, surge a oportunidade de uso dos públicos educacionais, uma área relativamente nova com potencial para auxiliar o desenvolvimento educacional [Bandeira et al. 2015]. Contudo, as fontes de dados educacionais no Brasil são publicadas separadamente sem permitir a ligação/navegação entre conjuntos de dados heterogêneos [Adeodato et al. 2014]. Além disso, cada fonte de dados tende a possuir um vocabulário próprio, onde um mesmo conceito pode ser representado por diferentes termos entre bases distintas.

Outra desvantagem é que os dados também não são publicados em formato padronizado, logo o acesso às informações pode advir de páginas web ou em formatos como: "*.xls*", "*.csv*", "*.pdf*" entre outros.

DOI: 10.5753/cbie.sbie.2018.1911 

Neste trabalho foram utilizadas tecnologias de *Linked Data* (em português Dados ligados) [Heath and Bizer 2011] e *Semantic Web* (em português Web Semântica) [Berners-Lee et al. 2001]. A utilização dessas tecnologias permite a criação e publicação de uma visão semântica unificada, chamada *Linked Data Mashups (LDMs)*. Um LDM permite que sejam desenvolvidas novas aplicações, transformando e agregando dados de fontes heterogêneas [Hoang et al. 2014]. Logo, o objetivo do trabalho consiste em realizar um *mashup* entre dados de indicadores educacionais e de execução financeira do PDDE.

## 2. Trabalhos Relacionados

Na literatura, muitos dos trabalhos relacionados à dados no domínio da educação utilizam o conceito de Mineração de Dados Educacionais (MDE). Nesta perspectiva, foram selecionados os trabalhos com objetivos mais próximos às temáticas aqui apresentadas. [Ferreira 2015] utiliza MDE aplicada sobre Microdados do Censo Escolar da Educação Básica do ano de 2014 visando identificar a influência de características familiares e individuais na conclusão do Ensino Fundamental. Similarmente, o estudo realizado por [Carvalho et al. 2017] objetiva analisar perfis de instituições e estudantes com base em dados dos Censos do Ensino Superior e Básico utilizando MDE. [Rigo et al. 2012] também utilizaram MDE para realizar um estudo que resulta em contribuições para detecção de comportamentos relacionados à evasão. Já de forma distinta, utilizando *Linked Data* ,[Cabral et al. 2012] apresentam um processo de integração de dados do Exame Nacional do Ensino Médio (ENEM) realizado no ano de 2008, tendo como contribuição a publicação dos dados em formato ligado.

Entretanto, de modo geral os trabalhos encontrados não abrangem o uso de dados sobre rendimento, média de alunos por ano e de execução financeira do PDDE no Ensino Básico. Assim, uma análise de aspectos envolvendo os estudos semelhantes é apresentada na Tabela 1.

**Tabela 1. Trabalhos Relacionados.**

| | *Linked Data* | PDDE | Média de Alunos | Rendimento |
|---|---|---|---|---|
| [Cabral et al. 2012] | X | - | - | - |
| [Rigo et al. 2012] | - | - | - | X |
| [Ferreira 2015] | - | X | X | - |
| [Carvalho et al. 2017] | - | X | X | - |
| Este estudo | X | X | X | X |

**Fonte: Elaborado pelos autores, 2018.**

## 3. Metodologia

O processo de integração foi baseado no *framework* LDIF. O LDIF sugere o seguinte fluxo de execução: *i)* Extração das fontes de dados; *ii)* Exportação das visões e transformação dos dados; *iii)* Resolução da identidade através de links *owl:sameAs*; *iv)* Avaliação da qualidade e fusão dos dados e *v)* Saída dos dados [Schultz et al. 2012].

Os dados obtidos do portal de microdados do INEP foram os *datasets* "Média de alunos por turma em escola" dos três últimos anos e "Taxas de rendimento" dos dois últimos anos disponíveis. Além disso, no repositório de dados abertos do PDDE foram selecionados os *datasets* "EXEC_FINANC_ED_BAS" (execução de financiamento

educação básica) dos anos 2016 e 2017. Os *datasets* sobre média de alunos contém informações básicas sobre a escola e a quantidade média de alunos por turma e nível. Os *datasets* taxas de rendimentos além de conterem informações básicas sobre a escola, também possuem informações sobre as taxas de aprovação, reprovação e abandono por turma e nível. Por fim os *datasets* de execução de financiamento possuem informações sobre as unidades executoras responsáveis pelo repasse de verba, custeio de verba e capital repassado e o CNPJ das UXs (Unidades Executoras das Escolas).

Com exceção dos *datasets* sobre financiamentos que já estavam em *".csv"*, todos os demais foram disponibilizados no formato *".xls"*, sendo necessário uma etapa de conversão para *".csv"*, por conseguinte, realizando a padronização de campos com valores nulos e numéricos além da limpeza do cabeçalho e notas.O processo de transformação dos dados para as visões exportadas em *rdf* foi realizado da seguinte maneira: os *datasets* no formato *".csv"* já tratados foram importados para o banco de dados relacional PostgreSQL e foram mapeados utilizando o *R2RML* [W3C 2012], sendo posteriormente *triplificados* pela ferramenta *D2RQ* [Bizer and Seaborne 2004].

Os mapeamentos responsáveis pela geração das visões exportadas utilizaram um padrão de criação de *URIs* que garante que a maioria das entidades de um mesmo objeto, que possuam códigos de identificação único, sejam representadas por um único recurso *RDF*. No entanto algumas entidades não possuíam identificadores uniformes entre as bases. Nestes casos foi necessário realizar a etapa de descoberta e materialização dos *links owl:sameAs* com uso da ferramenta *Silk* [Volz et al. 2009], em especial entidades da classe Unidade Executora, que apesar de possuírem um código de identificação único, sendo este o *CNPJ*, em diferentes bases o mesmo encontrava-se formatado de diferentes maneiras. Além da fusão é necessário que seja feita a limpeza dos dados, pois todas as propriedades possuídas pelos recursos fundidos serão referenciadas para um único recurso, podendo gerar conflitos ou valores repetidos. Esta atividade utilizou a ferramenta *Sieve* [Mendes et al. 2012].

## 4. Resultados

Neste trabalho foi produzido e publicado um *Linked Data Mashup* sobre dados indicadores da educação brasileira e de execuções financeiras do PDDE. Para validar o *mashup* foram realizadas consultas *SPARQL* sob um *Triple Store Virtuoso*. Um exemplo de consulta para retornar o repasse fornecido pelo PDDE e as respectivas taxas de aprovação e abandono pode ser visualizado na Listagem 1.

```
PREFIX oed: <http://www.blind.sbie.br/ontology/educa/>
SELECT DISTINCT ?cod ?escola ?qt_alunos ?repasse ?ano ?nivel ?aprovacao
  ?abandono ?total_alunosNivel
WHERE {
     ?a a oed:Escola;
          oed:codigo ?cod;
          oed:nome ?escola;
          oed:anoLetivo ?anoR.
     ?anoR oed:ano ?anoi;
          oed:quantidadeAlunos ?qt_alunos;
          oed:recebe ?repasseR.
     ?repasseR oed:total ?repasse.

    OPTIONAL{
          ?anoR oed:turma/oed:nivel ?nivelR.
          ?nivelR oed:descricao ?nivel;
            oed:totalAlunos ?total_alunosNivel.
             OPTIONAL{
```

```
                ?nivelR oed:totalAprovacao ?aprovacao;
                oed:totalAbandono ?abandono
            }
    }
    BIND(str(?anoi) as ?ano) FILTER(?ano = "2016")
} ORDER BY DESC (?repasse ) LIMIT 100
```

**Listagem 1: Consulta SPARQL Relação entre repasse e aprovação.**

Os resultados para a listagem 1 com uma amostra de 100 registros mostram que as escolas que possuem uma menor quantidade de alunos, por conseguinte, apresentam maiores taxas de abandono, refletindo no repasse do PDDE. Tal situação pode vir eventualmente a impactar na decisão de pais e alunos durante a escolha de qual instituição de ensino o aluno irá efetuar sua matrícula. Outro exemplo de consulta mostra a distribuição de alunos por escolas associada a suas taxas de rendimento no ensino básico, podendo ser visto na Listagem 2.

```
PREFIX oed: <http://www.blind.sbie.br/ontology/educa/>
SELECT DISTINCT ?local (count(?esc) as ?quantidade_escolas)
   (sum(?quat_alun) as ?total_alunos)
   (avg(?quat_alun) as ?media_alunos_por_escola) ?nivel
   (avg(?aprovacaoNivel) as ?taxa_aprovacao_nivel_media)
   (avg(?abandonoNivel) as ?taxa_abandono_nivel_media)
   (avg(?alunosNivel) as ?alunos_nivel_media)
WHERE {
       ?esc a oed:Escola;
            oed:localizacao ?local;
            oed:anoLetivo/oed:quantidadeAlunos ?quat_alun;
            oed:anoLetivo/oed:turma/oed:nivel ?nive.
       ?nive oed:descricao ?nivel;
            oed:totalAprovacao ?aprovacaoNivel;
            oed:totalAlunos ?alunosNivel;
            oed:totalAbandono ?abandonoNivel.
} GROUP BY ?local ?nivel LIMIT 100
```

**Listagem 2: Consulta SPARQL Distribuição de alunos por escola.**

Como resposta, a listagem 2 mostra que a maior taxa de evasão escolar se encontra no ensino médio, além de que o número de escolas de ensino médio no meio rural é substancialmente menor que no meio urbano, embora no nível fundamental os números sejam próximos. Os *datasets* e demais arquivos do projeto estão disponíveis no link do *DataHub*[1]

## 5. Conclusão

O *mashup* desenvolvido neste estudo representa uma visão unificada de múltiplas fontes de dados e permite a descoberta de novas informações através de consultas *SPARQL* envolvendo os dados sobre execuções financeiras e indicadores educacionais. Como trabalhos futuros pretende-se i) ampliar a variedade de dados utilizados durante os anos; ii) englobar outros indicadores educacionais; e iii) fornecer uma aplicação de acesso web.

## Referências

Adeodato, P. J., Santos Filho, M. M., and Rodrigues, R. L. (2014). Predição de desempenho de escolas privadas usando o enem como indicador de qualidade escolar. In *Simpósio Brasileiro de Informática na Educação-SBIE*, volume 25, page 891.

Bandeira, J., Ávila, T., Alcantara, W., Sobrinho, A., Bittencourt, I. I., and Isotani, S. (2015). Dados abertos conectados para a educação. *Jornada de Atualização em Informática na Educação*, 4(1):47–69.

[1]https://datahub.io/linkeddatamashupeducacional/educa-o/v/2

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific american*, 284(5):34–43.

Bizer, C. and Seaborne, A. (2004). D2rq-treating non-rdf databases as virtual rdf graphs. In *Proceedings of the 3rd international semantic web conference (ISWC2004)*, volume 2004. Proceedings of ISWC2004.

Cabral, S. P., Beduschi, N. B., Zancanaro, A., Todesco, J. L., and Gauthier, F. A. O. (2012). Aplicando linked data na publicação de dados do enem. In *ONTOBRAS-MOST*, pages 176–181.

Carvalho, J., Cruz, L., and Gouveia, R. (2017). Descoberta de conhecimento com aprendizado de máquina supervisionado em dados abertos dos censos da educação básica e superior. In *Anais dos Workshops do CBIE*, volume 6, page 674.

Ferreira, G. (2015). Investigação acerca dos fatores determinantes para a conclusão do ensino fundamental utilizando mineração de dados educacionais no censo escolar da educação básica do inep 2014. In *Anais dos Workshops do CBIE*, volume 4, page 1034.

FNDE (2017). PDDE programa dinheiro direto na escola. `http://www.fnde.gov.br/programas/pdde/sobre-o-plano-ou-programa/sobre-o-pdde`. Accessed: 2018-09-18.

FNDE (2018). PDDE dados abertos. `http://www.fnde.gov.br/dadosabertos/dataset?tags=PDDE`. Accessed: 2018-09-18.

Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136.

Hoang, H. H., Cung, T. N.-P., Truong, D. K., Hwang, D., and Jung, J. J. (2014). Retracted: Semantic information integration with linked data mashups approaches. *International Journal of Distributed Sensor Networks*, 10(4):813875.

INEP (2018). Microdados microdados do inep. `http://portal.inep.gov.br/microdados`. Accessed: 2018-09-18.

Júnior, G. C., Nascimento, R., Alves, G., and Gouveia, R. (2017). Identificando correlações e outliers entre bases de dados educacionais. In *Anais dos Workshops do CBIE*, volume 6, page 694.

Mendes, P. N., Mühleisen, H., and Bizer, C. (2012). Sieve: linked data quality assessment and fusion. In *Proceedings of the Joint EDBT/ICDT Workshops*, pages 116–123. ACM.

Rigo, S. J., Cazella, S. C., and Cambruzzi, W. (2012). Minerando dados educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In *Anais do CBIEão*, pages 168–177.

Schultz, A., Matteini, A., Isele, R., Mendes, P. N., Bizer, C., and Becker, C. (2012). LDIF - A Framework for Large-Scale Linked Data Integration. In *21st WWW, Developers Track*, page to appear.

Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G. (2009). Silk-a link discovery framework for the web of data. *LDOW*, 538.

W3C (2012). R2RML w3c recommendation r2rml. `https://www.w3.org/TR/r2rml/`. Accessed: 2018-09-18.