Criação de um repositório integrado de dados da execução orçamentária brasileira proveniente de diferentes fontes de dados em um modelo dimensional Data Warehouse

José R. Beluzo[1], Gisele S. Craveiro[1]

[1]Escola de Ciências, Artes e Humanidades – Universidade de São Paulo (USP)
Av. Arlindo Béttio, 1000 - Ermelino Matarazzo – São Paulo – SP – Brasil

{jrbeluzo,giselesc}@usp.br

Abstract. *This paper presents a proposal for data and schemes integration of public budget execution (revenues and expenses) in a Data Warehouse from the data available in portals transparency. The prototype integration proposal aims better information transparency of federal entities to better subsides researches on public policy, political science and also auditing and citizen participation in public policy in Brazil.*

Resumo. *Este trabalho apresenta uma proposta para integração de dados e esquemas da execução do orçamento público brasileiro (receitas e despesas) em um Data Warehouse a partir dos dados disponibilizados em portais de transparência. O protótipo de integração proposto prevê melhor transparência de informações dos entes federativos para auxilio em pesquisas de políticas públicas, ciências políticas e também auditoria e participação cidadã nas políticas públicas no Brasil.*

1. Introdução

O Acesso à informação pública é essencial para a transparência das ações de governo e a transparência orçamentária é fator chave em maior *accountability* [1]. No Brasil existe um arcabouço legal que obriga a publicação dos registros orçamentários em tempo real em portais de transparência pública na rede mundial de computadores – a Internet. Este arcabouço é composto pela Lei de Responsabilidade Fiscal (Brasil, 2000), Lei Capiberibe (Brasil, 2009) que obriga a União, os estados e os municípios a divulgar seus gastos na internet e Lei de Acesso a Informação (Brasil, 2012) que regulamenta o direito constitucional do cidadão ao acesso a informações produzidas pelo Governo.

Desde então vemos surgir diversos aplicativos cívicos que fazem uso de *mashups* [2] para facilitar o entendimento e análise dos dados publicados pelo governo. Exemplos destas aplicações são: "Para onde foi meu dinheiro" [3], que ajuda o cidadão a

---

1 http://www.oecd.org/governance/budgeting/Best%20Practices%20Budget%20Transparency%20%20complete%20with%20cover%20page.pdf

2 http://dados.gov.br/faq/#q5

3 http://www.paraondefoiomeudinheiro.org.br/

monitorar a execução dos orçamentos; e o "Cuidando do meu Bairro" [4] que mostra em um mapa a localização dos gastos previstos ou realizados nos equipamentos públicos dentro do município de São Paulo.

Os exemplos citados anteriormente mostram o interesse e o potencial que o reuso dos dados orçamentários disponíveis na web apontam. Porém ainda existem grandes desafios para que aplicativos relativos à transparência orçamentária ofereçam análise de dados mais complexas. Por exemplo: um especialista em políticas públicas que deseje analisar qual o valor gasto pela Função "Educação" nas três esferas do poder público em uma determinada cidade, este necessitará acessar três conjuntos diferentes de *datasets*. Caso este deseje obter a mesma informação do Estado Brasileiro em sua totalidade, necessitará acessar *datasets* de 1 portal federal, 26 portais estaduais, 1 portal do distrito federal e 5570 portais de municípios do Brasil.

Além do problema de acesso a diferentes bases de dados, encontramos também o problema de padronização. A falta de padronização dos dados torna a sua utilização trabalhosa, visto que para gerar uma análise baseada em três portais de transparência (suponha ainda o exemplo do especialista em políticas públicas) este precisará entender o esquema de publicação dos três portais e também como estes dados estão relacionados com suas respectivas receitas e despesas. Se expandirmos a análise para o Estado Brasileiro, novamente teremos que estudar os 5599 portais para conseguir realizar tal análise.

Do ponto de vista computacional, também há dificuldades na manipulação destes *datasets*, visto que o formato ao qual estes são encontrados estão ou em planilhas de texto (arquivos CSV – *Comma-separated Value*) ou arquivos XML acessados por protocolo HTTP. Em ambos os casos há grande redundância de valores, o que torna o processamento destes dados custosos. A falta de metadados e a perda semântica do modelo Entidade-Relacionamento sobre o esquema de dados dos *datasets* também dificulta o entendimento correto para análises e interpretações realizadas sobre os dados.

Em 2007 um grupo de ativistas publicou um manifesto na web que determina oito princípios básicos [5] para que um dado governamental publicado seja considerado como dado aberto. Em 2013 (último ano de adequação à Lei Capiberibe) foi realizado um trabalho de análise (Craveiro et al, 2013) que coletou e analisou mais de 300 *datasets* publicados por 88 portais de transparência e também Tribunais de Contas. A análise baseada nos oito princípios e nas leis Capiberibe e Lei de Acesso a Informação detectou que grande parte dos portais não se enquadram nos padrões de dados abertos governamentais e sequer atendem às necessidades exigidas pela lei.

A falta de integração dos dados assim como a falta de padronização da publicação dos dados da execução orçamentária é um problema atual. No Brasil ainda não existe uma solução que resolva este problema nesta perspectiva. Este trabalho possui como foco demonstrar que é possível resolver este problema utilizando uma integração de dados em um sistema de Informação do tipo *Data Warehouse*.

---

[4] http://www.gpopai.usp.br/cuidando

[5] http://opengovdata.org/

Este trabalho está dividido da seguinte forma: na Seção 2 define-se quais são os objetivos gerais e específicos e o escopo do trabalho a ser realizado para solução do problema citado; na Seção 3 é informado quais são as áreas e de que forma este trabalho irá contribuir, assim como a delimitação do escopo a ser atingido; na Seção 4 é demonstrado o estágio que se encontra o trabalho e quais foram as dificuldades encontradas até o momento; e na Seção 5 discute-se os resultados que se deseja obter com este trabalho, as limitações encontradas e possíveis trabalhos futuros.

2. Objetivos da Pesquisa

Este trabalho tem como objetivo principal criar um repositório de dados integrados em um Sistema de Informação do tipo "*Data Warehouse*" a partir de diferentes fontes de dados heterogêneas em estruturas, sintaxe e formatos de dados referentes à execução orçamentária no Brasil (receitas e despesas) nas três esferas governamentais: Federal, Estadual e Municipal. Além disso, o trabalho contemplará:

- Limpeza de dados – existe além do problema da integração e padronização o erro humano, ao qual consiste em erros de digitação de valores e falta de valores;
- Proveniência de dados – informações das fontes iniciais dos dados, assim como as transformações realizadas nos dados até que estes estejam no formato adequado para o novo esquema;
- Reconstrução semântica – informações contextualizando o relacionamento existente entre os campos de dados de cada fonte, além da relação existente entre os campos de cada ente federativo.

3. Contribuições Esperadas

Esse trabalho pretende contribuir na coleção, limpeza, organização dos dados de execução orçamentária no Brasil construindo e povoando um repositório de dados que permitirá a realização de consultas acerca do gasto realizado para implementar diferentes políticas públicas nos diversos níveis de governo.

Pretende-se que esse grande armazém de dados possa auxiliar pesquisadores das mais diferentes áreas e que necessitam se debruçar sobre a massa de dados orçamentários que hoje é disponibilizada de forma heterogênea em relação à semântica, granularidade, qualidade e de formato.

Além disso, espera-se que esse repositório possa futuramente ser republicado na web e assim oferecer um canal de acesso à população aos dados orçamentários de forma concentrada e padronizada.

Inicialmente, como prova de conceito, o escopo do trabalho será delimitado à execução orçamentária em nível Federal (Portal da Transparência do Governo Federal [6]); Estadual (Portal de Transparência da Secretária da Fazenda do Estado de São Paulo [7]) e Municipal (Portal do Cidadão – Tribunal de Contas do Estado de SP [8] e o Portal da

---

[6] http://www.portaltransparencia.gov.br/

[7] http://www.fazenda.sp.gov.br/download/default.shtm

Transparência do Município de São Paulo [9]). A partir deste protótipo, deseja-se mostrar a viabilidade do projeto e que a sua expansão para os demais entes federativos (trabalhos futuros) será útil e possibilitará melhores análises, pois possibilitará a combinação de valores nos três entes federativos para todos estados e municípios.

4. Resultados já alcançados

O primeiro passo executado neste trabalho foi realização de uma revisão sistemática em bases de dados nacionais (busca realizada em BDBComp – Biblioteca Brasileira de computação – e Teses USP pelas palavras-chave "*Data Warehouse*" ou "Integração de Dados"; e Google *Schoolar* por "*Data Warehouse* Orçamento Público") e internacionais (busca realizada nas bases de dados IEEE, ACM e Google *Schoolar* pelas palavras-chave "*Open Government Data*", "*Open Data*" e "*government data integration*") sobre os temas relacionados ao escopo da pesquisa.

Dentre os trabalhos retornados, foi realizada uma seleção baseada em títulos e resumos dos trabalhos que possuíam conteúdo relacionado ao assunto. Dentre estes, na busca internacional três trabalhos tiveram destaque por possuírem um tema relacionado ao assunto de pesquisa: Midas, GovWild e EnAKTing.

Os projetos Midas (Sala et al, 2010) e GovWild (Böhm et al, 2012) criaram uma arquitetura de centralização de dados orçamentários provenientes de vários portais de transparência. Por meio de uma ferramenta de buscas na WEB, é possível retornar e exportar dados das bases de dados de forma integrada. Este trabalho possui características semelhantes aos trabalhos citados. A proposta deste trabalho é a integração dos três níveis federativos no Brasil, diferencial com relação aos trabalhos citados, visto que este integram apenas dados a nível federal em seus trabalhos.

EnAKTing (Shadbolt et al, 2012) demonstra em seu trabalho a possibilidade de descoberta de conhecimento combinando fontes de dados de áreas diferentes. Para isso, foi utilizado um conceito computacional denominado LDW (*Linked Data* WEB) baseado nos princípios definidos pelo *Linked Data* [10], que afirma que é possível gerar um relacionamento dos dados que estão publicados (no formato proposto, RDF – *Resource Description Framework*), bastando apenas que se crie um descritor do relacionamento existente entre estes nos conjuntos de dados. EnAKTing não está diretamente relacionado com o trabalho aqui proposto, porém com trabalhos futuros, já que a possibilidade de integração de dados integrados do orçamento público com outros conjuntos de dados externos dependem de uma base de dados integrada do orçamento, demonstrando mais uma vez a necessidade da criação do repositório integrado.

Referente a busca nacional, existem trabalhos relacionados a padronização da publicação dos dados (Santana, 2013) governamentais do orçamento público por meio de uma taxonomia, que padroniza a forma de publicação dos dados. Outros trabalhos governamentais preveem a integração dos dados na esfera governamental Federal (Santos, 2011) – SIC (Sistema de Informação de Custos do Governo Federal), porém

---

[8] http://www.portaldocidadao.tce.sp.gov.br/

[9] http://transparencia.prefeitura.sp.gov.br/Paginas/home.aspx

[10] http://www.w3.org/designissues/linkeddata.html

com foco voltado a apenas gestores governamentais (Santos, 2011). Este trabalho prevê uma integração de dados nas três esferas governamentais: Federal, Estadual e Municipal, além de possibilitar o acesso a informação amplo para tanto gestores, pesquisadores e cidadão em geral.

Após a etapa de revisão bibliográfica, foi realizado o levantamento das bases de dados disponíveis para o processo de integração assim como o processo de extração dos *datasets*. Os dados a nível Federal foram extraídos do Portal da Transparência do Governo Federal [11]; A nível Estadual os dados foram extraídos do Portal da Secretária da Fazenda do Governo de São Paulo [12]; a nível Municipal os dados foram extraídos do Portal do Tribunal de Contas do Estado de São Paulo [13] pois este órgão concentra os dados referentes às despesas e receitas de todos os municípios do Estado de São Paulo, exceto do munícipio de São Paulo que possui seu próprio Tribunal de Contas. Para o município de São Paulo foi utilizado os *datasets* disponibilizados no Portal da Transparência do município de São Paulo [14].

A etapa seguinte a coleta de dados é a reconstrução semântica dos dados utilizados baseados nas regras do Manual Técnico do Orçamento Público (Brasil, 2013), etapa atual do projeto. Esta reconstrução semântica consiste em criar, a partir dos requisitos do MTO o diagrama entidade-relacionamento de cada esfera orçamentária, e, a partir destes diagramas, criar a intersecção dos modelos para gerar o modelo integrador. Esta reconstrução semântica permite melhor republicação destes dados em cada nível federativo e também auxilia a republicação do modelo integrado em trabalhos futuros. A Figura 1 ilustra as etapas ao qual compõem o projeto:

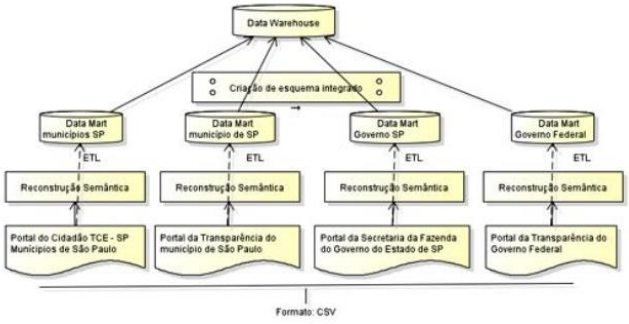

Figura 1. Processo de Integração de Dados Orçamentários nas 3 esferas

5. Conclusão

Este trabalho demonstra uma problemática a ser resolvida com relação a integração de dados da execução orçamentária brasileira referente aos três entes federativos e criação de um *Data Warehouse* integrando estes dados. A resolução deste problema propõe grandes melhorias na construção do conhecimento referente ao

[11] http://www.portaldatransparencia.gov.br/downloads/

[12] http://www.fazenda.sp.gov.br/download/default.shtm

[13] http://www.portaldocidadao.tce.sp.gov.br/downloads-e-api

[14] http://transparencia.prefeitura.sp.gov.br/acesso-a-informacao/Paginas/default.aspx

Orçamento Público, seja este para um especialista da área (Gestores Públicos e Pesquisadores de Políticas Públicas, por exemplo), seja este o cidadão comum, que irá ser capaz de melhor entender o orçamento e participar ativamente nas tomadas de decisões públicas e também ajudar no monitoramento e combate à corrupção no país. O escopo deste trabalho se restringiu a um ente estadual e seus entes municipais para se obter uma prova de conceito, além do ente federal. A integração dos outros entes federativos poderá ser inserida no modelo na medida em que este traga retornos positivos, seguindo-se da mesma metodologia proposta neste trabalho.

6. Referências

Araújo, Paulo Sérgio Sabino. A Tecnologia de Informação como ferramenta de Transparência Orçamentária: Evolução dos Sistemas Orçamentários e o Desafio da Integração Governamental. 2008.

Bernadino, Jorge. Open Source Business Intelligence Platforms for Engineering Education. SEFI Annual Conference, 2011.

Böhm, Christoph; Freitag, Markus; Heise1, Arvid; Lehmann, Claudia; Mascher, Andrina; Naumann, Felix. GovWILD: integrating open government data for transparency. Proceedings of the 21st international conference companion on World Wide Web, 2012.

BRASIL; Ministério do Planejamento, Orçamento e Gestão. Manual técnico de orçamento MTO - Versão 2014. 2013.

BRASIL, Lei Complementar no 131. Constituição da República Federativa do Brasil. 2009.

BRASIL, Lei no l2.527. Constituição da República Federativa do Brasil. 2011.

BRASIL. Constituição da República Federativa do Brasil. 1998.

Craveiro, Gisele da Silva; Santana, Marcelo Tavares; Albuquerque, João Porto. Assessing Open Government Budgetary Data in Brazil. ICDS 2013, The Seventh International Conference on Digital Society, 2013.

Sala, Antonio; Lin, Calvin; Ho, Howard. Midas for government: Integration of government spending data on Hadoop. In Proc. of the Int. WS on New Trends in Information Integration (NTII), 2010.

Santana, Marcelo Tavares. Uma proposta de publicação de dados do orçamento público na Web. Dissertação apresentada como parte da avaliação do programa de mestrado em Sistemas de Informação da Escola de Artes, Ciências e Humanidades EACH. 2013.

Santos, Welington Vitor. Sistema de Informação de Custos do Governo Federal: Modelo Conceitual, Solução Tecnológica e Gestão do Sistema. Secretaria do Tesouro Nacional. 2011.

Shadbolt, Nigel; O'Hara, Kieron; Berners-Lee, Tim; Gibbins, Nicholas; Glaser, Hugh; Hall, Wendy; Schraefel, M.C. Linked Open Government Data: Lessons from Data.gov.uk. Intelligent Systems, IEEE, 2012.