Visualização de Informação Aplicada ao Portal de Dados Abertos do IFSP

Fabio S. Oliveira [1], **Gustavo Bartz Guedes**[1]

[1]Campus Hortolândia – Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) 13183-250 – Hortolândia – SP – Brasil

fasantos4@gmail.com, gubartz@ifsp.edu.br

Abstract. *Brazilian government provides public data in a website called transparency portal. The huge volume of data becomes an issue to information comprehension. The objective of this work is to apply information visualization to facilitate the comprehension of the transparency portal data focusing on the Federal Institute of São Paulo (IFSP). The data have been consolidated using Pentaho, a data integration tool and Tibco Spotfire as the visualization tool to create visual representations.*

Resumo. *O portal da transparência da união disponibiliza dados do Governo federal, porém o volume de dados é demasiado para que as informações sejam compreendidas sem tratamento prévio. O objetivo deste trabalho é aplicar visualização de informação ao portal da transparência da união com foco nos dados do Instituto Federal de São Paulo. Para alcançar o objetivo o Pentaho foi utilizado para processar os dados e o Tibco Spotfire para criar visualizações.*

1. Introdução

Visualização de informação (InfoVis) é o estudo da transformação de dados abstratos em representações visuais com o propósito de aumentar sua compreensão e auxiliar na tomada de decisões [Card et al. 1999].

Aplicar visualização de informação para criar visões de dados do setor público amplia a transparência para os cidadãos e facilita a tomada de decisões para gestores do serviço público.

A Lei Complementar 131, de 27 de maio de 2009, determina que a União Federal, Estados, Municípios e o Distrito Federal devem disponibilizar todos os dados de despesas e receitas [BRASIL 2009].

O Portal da Transparência da União segue as determinações da LC 131/2009, disponibilizando os dados relacionados às despesas e o cadastro dos servidores públicos dos órgãos federais [BRASIL 2018]. Os arquivos disponibilizados são grandes e causam sobrecarga de dados dificultando a compreensão das informações tanto pelos cidadãos quanto pelos gestores do setor público.

Em agosto de 2017, com o objetivo de cumprir a Lei de Acesso à Informação 12.527/11 (LAI), foi aprovado o Portal da Transparência do IFSP denominado Portal de Dados Abertos (PDA). O portal limita a origem dos dados apenas a uma instituição e, com isso, melhora o acesso às informações. Entretanto, são disponibilizados no formato *Comma-Separated Values* (CSV), um arquivo de texto no qual os valores são separados

por vírgulas. Além disso, por se tratar de um grande volume de dados a leitura é dificultosa, mesmo que se use um editor de planilhas eletrônicas para visualizá-los em formato de tabela.

Portanto, este trabalho aplica a Visualização de Informação nos dados do Portal de Dados Abertos do IFSP com o objetivo de tratar a sobrecarga de dados transpondo-os para representações visuais (visões).

2. Referencial Teórico

Esta seção apresenta o referencial teórico que serviu de base para a realização deste trabalho de conclusão de curso.

2.1. Visualização de Informação

Dados são os elementos básicos que tomados isoladamente não permitem o entendimento de um contexto. Informação são conjuntos de dados dentro de um mesmo contexto que podem resultar em conhecimento.

A visualização de informação (InfoVis) utiliza representações visuais e interativas com o propósito de ampliar a cognição. Portanto, são úteis quando a atividade de avaliação ultrapassa a capacidade humana de análise e interpretação [Silva 2006].

A Visualização de Informação está inserida no dia a dia, visto que diariamente são utilizadas representações visuais para facilitar a compreensão de dados, como por exemplo os mapas meteorológicos utilizados em telejornais para informar a previsão do tempo, que são criados a partir de cálculos complexos e análise de dados da superfície, oceanos e atmosfera. Apesar desses dados demonstrarem a previsão do tempo, o grande volume não permite que os telespectadores compreendam em poucos minutos a previsão para sua região, por esse motivo são criadas visualizações para que seja transmitida de maneira clara. Este é um exemplo da aplicação da visualização de informação para resolver um problema de sobrecarga de dados.

A visão humana é treinada para identificar padrões. esta capacidade é demonstrada na Figura 1. Ao olhar para a imagem a visão imediatamente detecta a forma que está fora do padrão e muda o foco. Por essa razão, utilizar representações visuais ajuda na tomada de decisões, melhora a compreensão da informação e contribui para gerar conhecimento [Nascimento and Ferreira 2005].

Figura 1. Utilizando a visão para encontrar padrões.

Um outro exemplo se refere a verificação da evolução de um aluno durante um curso de graduação cujas médias são apresentadas na Figura 2 em formato de tabela.

5,00
5,70
7,00
7,50
6,30
7,00
7,00
8,00
8,50
9,00

Figura 2. Exemplo média geral do aluno por semestre.

Com o uso de InfoVis as médias do aluno foram transpostas para uma visão, um gráfico de barras, conforme mostra a Figura 3. O gráfico simplifica de maneira clara a evolução do aluno durante o curso, pois possibilita identificar de imediato o semestre em que o aluno estava acima da media e facilita a comparação com outros períodos.

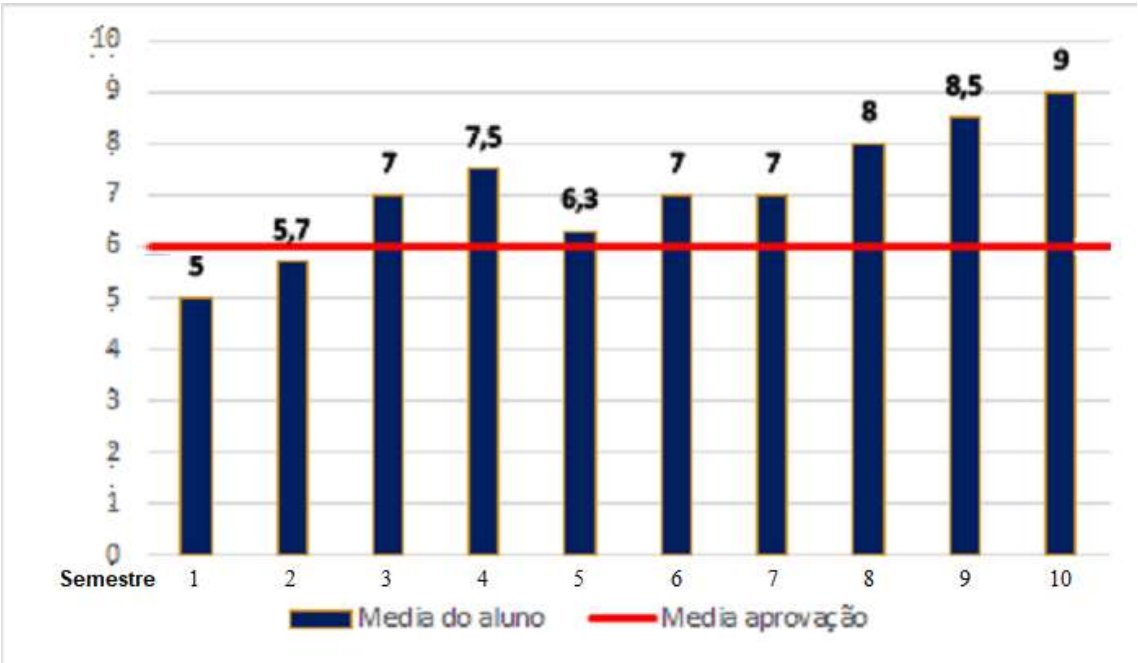

Figura 3. Exemplo de InfoVIs evolução durante o curso.

Segundo o modelo de visualizações de informação de [Card et al. 1999], mostrado na Figura 4, a criação de visualizações de informação inicia-se por meio da interação com os dados brutos, valores sem formatação ou tratamento prévio. Na transformação dos dados, os dados brutos são processados a fim de estruturá-los e eliminar redundâncias, dados incompletos e filtrar os relevantes.

Os dados podem ser estruturados em tabelas, nas quais cada linha ou registro representam um dado e as colunas representam os atributos. Na segunda etapa do modelo, os atributos dos dados são trabalhados para o mapeamento visual, em que as tabelas são transformadas em representações ou estruturas visuais, que por sua vez podem ser refinadas na transformação da visão, possibilitando a compreensão de uma informação com menos esforço. Por fim, são geradas visões.

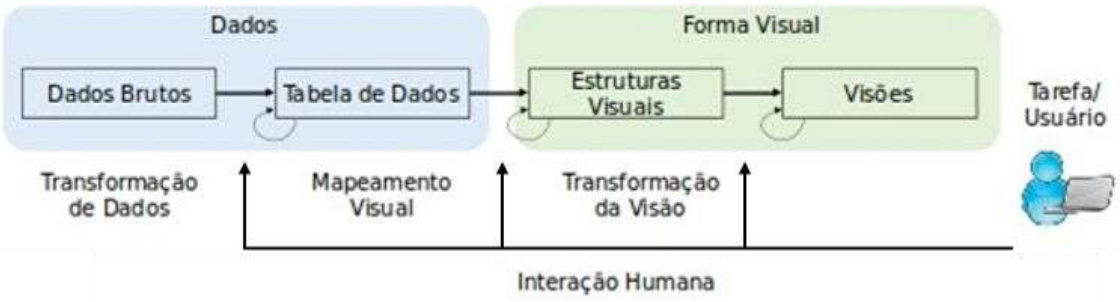

Figura 4. Modelo de referência [Card et al. 1999].

2.2. Técnicas de visualização de informação

Durante a criação de uma visualização, os dados são mapeados para atributos como cor, posição, área, tamanho e, para o caso de textos, é possível utilizar a fonte.

Visualizações podem ser interativas, possibilitando a manipulação da visão para dar ênfase em alguns elementos ou mudar a perspectiva para alterar o volume de detalhes da visão. A seguir são exemplificadas algumas visões utilizadas em InfoVis.

Tag Cloud ou Word Cloud, traduzido como nuvem de palavras, é uma visão utilizada para dados no formato de texto. A frequência das palavras é o atributo considerado para determinar a cor, traço e tamanho da fonte a fim de demonstrar sua importância no texto.

A Figura 5 é um exemplo de Tag Cloud criada a partir do discurso de posse do Presidente Barack Obama feito em Janeiro de 2008. A visualização foi criada no TagCrowd (Steinbock, 2018). Na visão do discurso do ex-presidente Barack Obama as palavras Nação, América, Geração, Pessoas e Trabalho são as 5 palavras-chave do discurso, por esse motivo são maiores e possuem um tom mais escuro de azul.

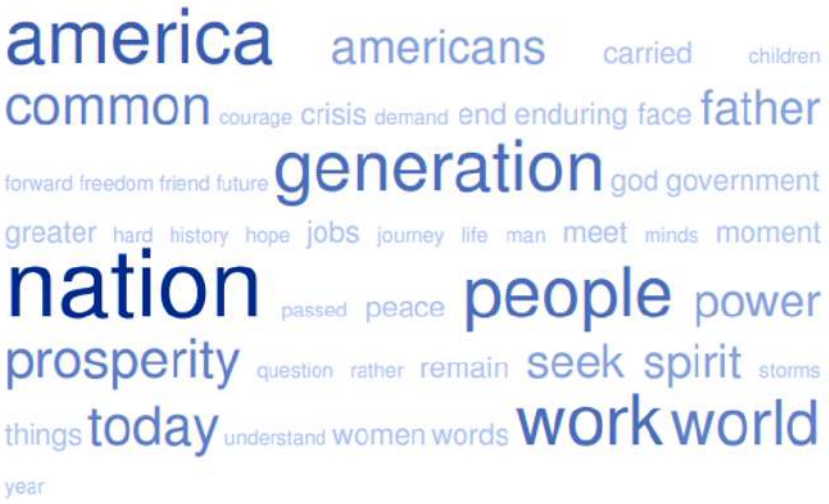

Figura 5. Tag Cloud do discurso de posse do Presidente Obama em 2008.

Treemap (árvore estruturada) representa hierarquias utilizando o conceito de árvore em que os galhos e folhas são representados por retângulos e a classificação de importância é determinada pelos atributos de cor, área e posição.

A Figura 6 é um Treemap criado pelos organizadores do Observatório de Complexidade Econômica (OEC, sigla em inglês), que utiliza os dados das exportações do Brasil em 2016 para criar a visão dos maiores importadores de produtos brasileiros.

Na visualização, as cores representam o continente e cada país é representado por um retângulo. A área do retângulo representa a porcentagem de produtos comprados do Brasil pelo país. Sendo assim os retângulos da China e dos Estados Unidos são os que possuem a maior área e consequentemente são os maiores importadores de produtos brasileiros. Sendo a China o maior importador, sua posição é a primeira da visão.

Figura 6. Treemap de importação de produtos brasileiros com base no país e continente no ano de 2016.

2.3. Ferramentas de Visualização de Informação

Existem ferramentas de InfoVis para dados em formato de texto, números e tabelas que implementam o modelo de referência de [Card et al. 1999] na geração de visões. A seguir serão descritas algumas ferramentas de InfoVis.

Gephi é uma ferramenta de código aberto desenvolvido em Java por estudantes da *University of Technology of Compiègne* (UTC - França) e atualmente mantida pela *Gephi Consortium* [Gephi Consortium 2017]. A representação dos dados é feita em grafos que são visões compostas por nós que representam um determinado dado ou conjunto, e arestas cujo objetivo é mostrar relacionamentos.

Tibco Spotfire é uma ferramenta para fontes de dados heterogêneas. Permite criar visualizações com gráficos de barras, gráfico de setores, Treemap, gráficos de dispersão, entre outras. É uma plataforma paga, mas disponibiliza versão gratuita com funcionalidades limitadas [TIBCO Software Inc 2018].

Watson Analytics, ferramenta de visualização que utiliza o supercomputador Watson para criar representações visuais. A versão gratuita da ferramenta possui funcionalidades limitadas para leitura de dados e não possui versão *desktop* [IBM 2018b].

2.4. Ferramentas de integração de dados

Ferramentas de integração de dados possuem a funcionalidade de processar e uniformizar dados estruturados e não estruturados de fontes heterogêneas. Algumas dessas ferramentas são descritas a seguir.

IBM DataStage, plataforma de integração de dados da IBM, possui integração com Watson Analytics e outras ferramentas. Um aspecto a se considerar é a latência para carregar os dados, uma vez que a plataforma é para internet, além disso as funcionalidades são limitadas na versão gratuita [IBM 2018a].

Pentaho, ferramenta inicialmente desenvolvida como código aberto, cujos direitos foram adquiridos pela Hitachi Vantara, que manteve uma versão gratuita chamada Pentaho Data Integration (PDI). O PDI é multiplataforma e permite a conexão com diferentes sistemas de gerenciamento de bancos de dados [Hitachi Vantara 2018].

2.5. Portal da Transparência

Em 2004, a Lei Complementar 131, conhecida como Lei da Transparência, foi sancionada e determinou que estados, municípios e a União deveriam publicar as informações sobre recursos públicos na internet. No mês de novembro do mesmo ano a Controladoria Geral da União (CGU) implantou o Portal da Transparência da União, um sítio que disponibiliza dados referentes aos gastos públicos em formato HTML e mantém alguns dos dados disponíveis no formato CSV.

Em agosto de 2017 o Portal da Transparência do IFSP foi lançado com o nome Portal de Dados Abertos com o objetivo de cumprir a Lei de Acesso à Informação. Os dados são disponibilizados em formato CSV e são relacionados às despesas, servidores e cursos do IFSP. Diferentemente do Portal da Transparência da União os arquivos do PDA são menores e divididos em categorias.

3. Trabalhos Correlatos

Nesta seção são apresentados trabalhos que utilizam o conceito de visualização de informação aplicados em dados públicos.

No trabalho "Visualização de Informação Aplicada à Legislação Penal Brasileira" técnicas de InfoVis foram aplicadas no banco de dados do Sispenas com o objetivo de facilitar a compreensão do código penal [Oliveira and Guedes 2015]. O Sispenas é um projeto criado por estudantes da faculdade de direito da Fundação Getúlio Vargas (FGV) e subsidiado pelo projeto Pensando Direito da Secretaria de Assuntos Legislativos, o Sispenas mantém o código penal brasileiro em um banco de dados relacional. Neste trabalho foram criadas representações visuais capazes de encontrar relacionamentos entre artigos do código penal e os critérios de aplicação de penas.

A Figura 7 é uma das visualizações criadas pelos autores, na qual a técnica Tag Cloud foi utilizada para destacar as palavras-chave do texto com o objetivo de aumentar a compreensão do capítulo da lei em destaque.

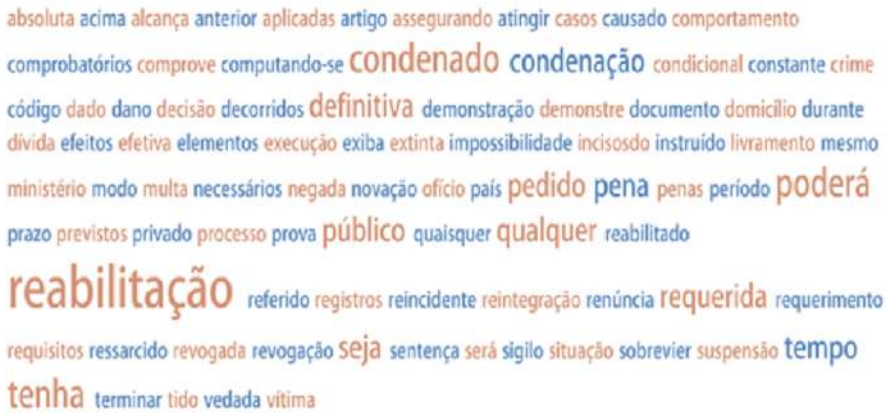

Figura 7. TagCloud do texto contido no Capítulo VII do decreto lei 2.848 [Oliveira and Guedes 2015].

A criação de visualizações para os dados públicos amplia a transparência dos dados do governo. Como consequência, toda a população tem acesso às informações de maneira clara, o que facilita sua compreensão. Assim a confiança e satisfação dos cidadãos em relação ao governo cresce [Papaloi and Gouscos 2013].

O trabalho de [Paula et al. 2011] utiliza dados do programa Bolsa Família para gerar visões que representam a quantidade de bolsas concedidas por região utilizando gráficos de barras e um mapa do Brasil para representar as informações por estados e regiões brasileiras. O mapa do Brasil com distorções da Figura 9 é a opção que apresenta as informações de maneira mais intuitiva, pois representa os estados com as cores destacando a quantidade de benefícios oferecidos.

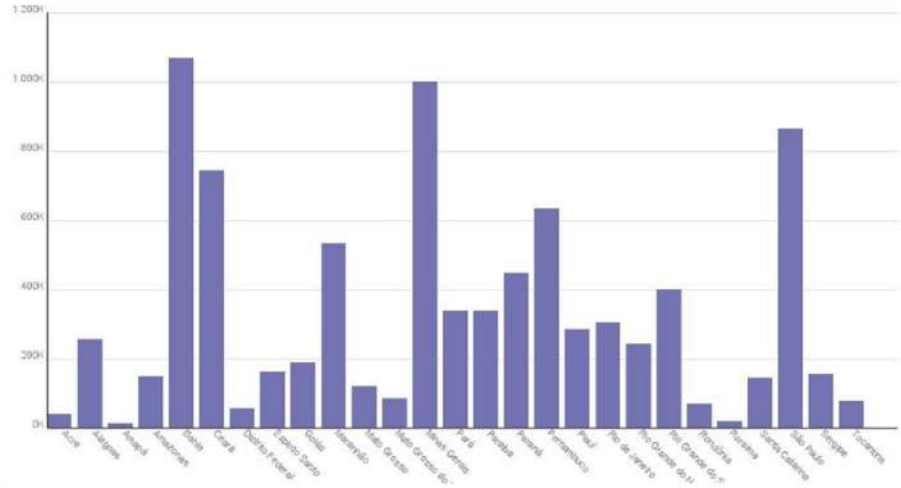

Figura 8. Gráfico de barras representando a quantidade de pessoas que recebiam o bolsa família por estado em 2005 [Paula et al. 2011]

.

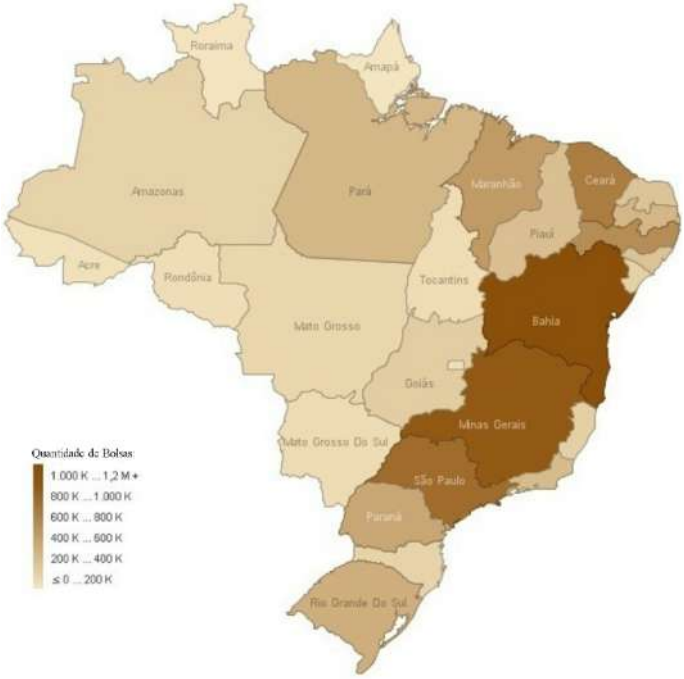

Figura 9. Visualização do mapa do Brasil representando a quantidade de pessoas que recebem o bolsa família por estado em 2005 [Paula et al. 2011]

.

Todos os trabalhos apresentados aplicam visualização de informação para melhorar a compreensão dos dados pela população, porém a coleta e transformação dos dados é manual.

4. Metodologia

Inicialmente foi feita uma análise dos dados do Portal da Transparência para decidir qual o conjunto de dados a ser trabalhado. Foi decidido por aplicar Visualização de Informação nos dados do Instituto Federal de São Paulo, disponibilizados no PDA.

Foi realizada uma pesquisa por softwares de InfoVis e softwares de integração de dados tendo como critério de avaliação a quantidade de funcionalidades disponíveis na versão gratuita. Os softwares escolhidos foram Pentaho (PDI) para integração de dados e Tibco Spotfire como ferramenta de InfoVis.

Por fim, foi feita uma inspeção da estrutura do HTML do PDA a fim de desenvolver um *webcrawler* na linguagem JavaScript. Optou-se por armazenar os dados em um banco de dados relacional utilizando o sistema de gerenciamento de bancos de dados PostgreSQL por possuir compatibilidade com o Tibco Spotfire.

As próximas seções apresentam as etapas de desenvolvimento do trabalho.

5. Desenvolvimento

Nesta seção é apresentado o desenvolvimento do trabalho representado pela Figura 10.

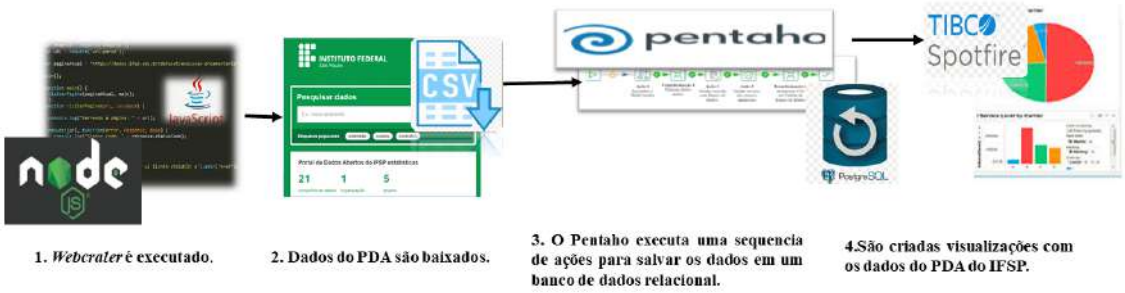

Figura 10. Processo de coleta e transformação dos Dados.

5.1. Coleta dos dados

Neste trabalho foram coletados arquivos do PDA referentes aos servidores, cursos e execução orçamentária relativos ao período de 2009 a 2017 contendo dados de todos os *campi* do IFSP. A coleta dos dados foi automatizada por meio do *webcrawler*, que analisa o código-fonte em HTML do PDA e faz o *download* dos arquivos de acordo com o nome do conjunto de dados.

O Pentaho possui a funcionalidade de enviar comandos ao sistema operacional. Esta funcionalidade foi utilizada para iniciar o NodeJS e executar o *webcrawler*. O processo é representado pelos itens 1 e 2 da Figura 10.

5.2. Integração dos dados

A Figura 11 mostra o processo de integração dos dados, que consiste em cinco etapas descritas a seguir.

Figura 11. Processo de integração dos dados.

1. A Ação 1 realiza a coleta dos dados do PDA por meio do *webcrawler* (seção 5.1).
2. A Transformação 1 faz a leitura do arquivo CSV e o Pentaho identifica automaticamente o tipo de dado de cada campo, verifica a existência de valores faltantes e remove os registros duplicados do arquivo. Essa é a etapa de maior criticidade, pois uma falha pode gerar problemas de integridade no banco de dados. A lógica garante que não existam valores faltantes ao excluir os registros que possuem todas as colunas vazias, e em caso de registros duplicados mantém apenas a primeira ocorrência encontrada.
3. A Ação 2 verifica a existência do banco de dados de destino. Caso o banco de dados não exista, o processo é finalizado com uma mensagem de erro. Essa ação pode ser reiniciada após a criação do banco de dados.
4. A Ação 3 verifica se não existem registros com erro para que os dados sejam mapeados com a formatação correta, pois existe a possibilidade de haver registros com separadores diferentes da vírgula ou de não existir separador entre dois campos por erros durante a criação do arquivo CSV. Para esses dois casos, o Pentaho, por padrão, mescla os campos causando erros de sintaxe no momento de incluir o registro na tabela. Pelo fato do arquivo CSV ser extenso não é possível fazer esta análise previamente para corrigir o arquivo ou indicar ao Pentaho que devem ser considerados outros separadores além da virgula.
5. A Transformação 2 é a etapa de mapeamento do CSV para tabelas. Durante esta transformação o Pentaho cria e insere os dados nas tabelas do banco de dados destino.

6. Resultados

A ferramenta utilizada neste trabalho para aplicar as técnicas de visualização de informação foi o Tibco, pois disponibiliza maior quantidade de funcionalidades em sua versão gratuita.

A Figura 12 é um Treemap que compara a quantidade de professores e técnicos-administrativos no IFSP por *campus* em abril de 2018. Na visualização, a área, cor e posição de cada nó são definidas pela quantidade de servidores. Sendo assim, o nó com maior peso é posicionado no canto superior esquerdo e representa o *campus* com mais servidores. Já a cor representa os cargos, sendo que as de tonalidade mais escura são os cargos com maior número de servidores.

A análise do Treemap evidencia que a maior parte dos servidores são professores e que o *campus* São Paulo possui o maior número de servidores, o que se justifica pelo fato de ser o mais antigo, fundado em 1.909.

Figura 12. Relação de servidores por cargo e *campus*.

O Treemap da Figura 13 tem como referência os dados de abril de 2018 dos cursos oferecidos pelo IFSP por *campus*. Cada nó representa um *campus* e sua área é definida pela soma de cursos oferecidos. Os nós internos representam os cursos oferecidos e seu tamanho a quantidade de turmas. Por fim, a cor diferencia os tipos dos cursos (EJA, Gradução, Pós-Graduação e Técnico). A análise do Treemap evidencia que o *campus* São Paulo possui a maior variedade de cursos, justificando que este seja o *campus* com maior quantidade de funcionários. O nó IFSP, logo a baixo do *campus* São Paulo, possui mais funcionários com o cargo de assistente em administração, o que é esperado, pois o nó representa a reitoria do IFSP.

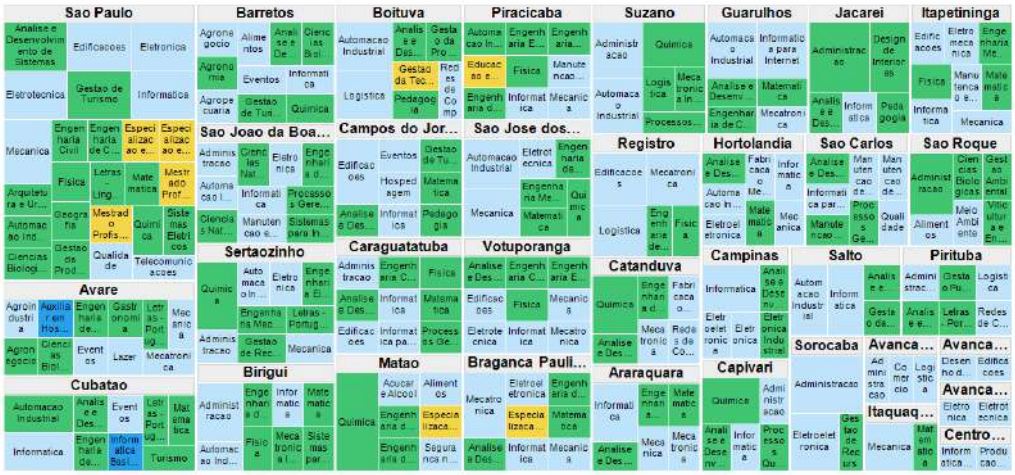

Figura 13. Relação de cursos por campus.

O Tibco permite a interação do usuário com a visualização, como por exemplo, ampliar um nó do Treemap por meio da ação de clique. A Figura 14 é a visão ampliada do nó Hortolândia. Essa visão responde rapidamente quais cursos o campus oferece e quais os tipos de curso.

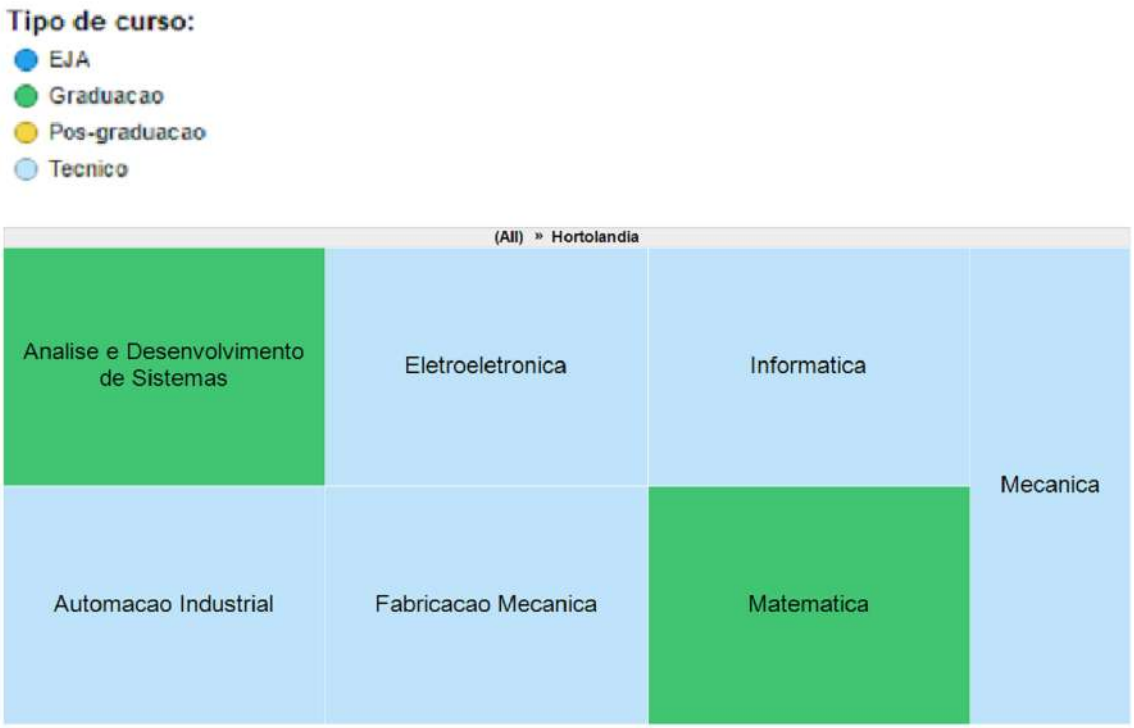

Figura 14. Visão dos cursos oferecidos no *campus* Hortolândia.

O gráfico de barras comparativo da Figura 15 mostra um comparativo do total de servidores em relação à execução orçamentária de alguns *campi* do IFSP, em que o tamanho das barras representa o orçamento do *campus* do ano de 2017. As barras estão ordenadas de maneira crescente, de acordo com a quantidade de funcionários indicada a baixo do nome do *campus*, e a cor diferencia os *campi* com menos de 100 funcionários. A visualização infere que o tamanho do *campus* não é proporcional ao orçamento, tomando como exemplo os *campi* São Carlos e Hortolândia. Mesmo com uma diferença pequena na quantidade de servidores, Hortolândia com 113 e São Carlos com 111, o *campus* São Carlos apresentou um orçamento maior que o campus Hortolândia, demonstrando que o orçamento não é dependente direto da quantidade de funcionários.

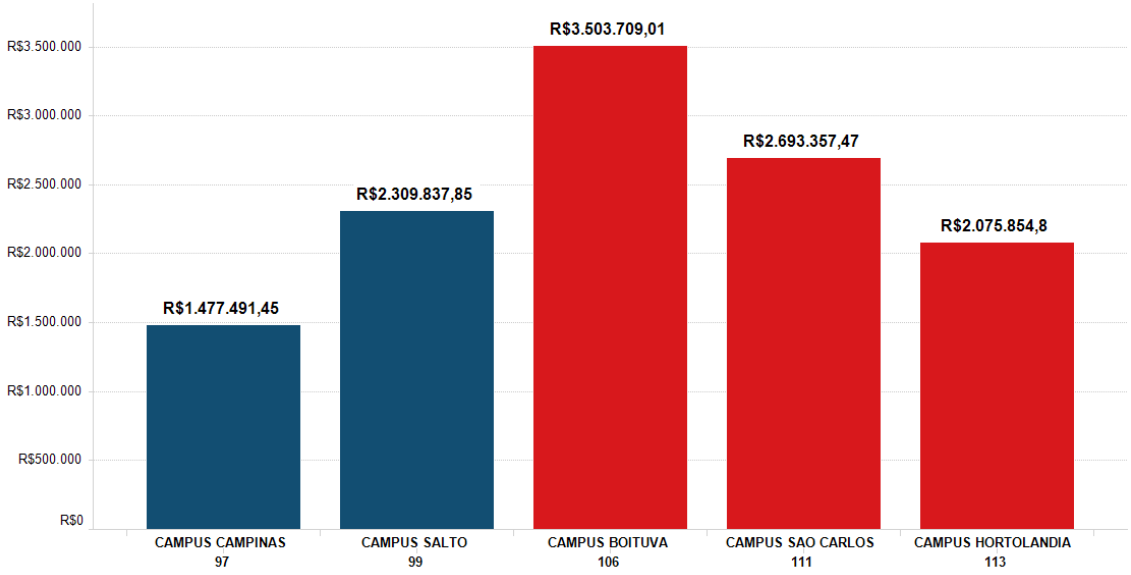

Figura 15. Comparação entre orçamento e servidores nos campi Boituva, Campinas, Hortolândia, Piracicaba, Salto e São Carlos durante o ano de 2017.

A visualização da Figura 16 mostra o orçamento para bolsas de estudo do *campus* Hortolândia entre os anos de 2015 e 2017.

A visualização mostra que houve uma queda no orçamento para bolsas de estudo no ano de 2016. Neste período o país passava por uma fase de instabilidade econômica que pode ter afetado o orçamento disponível para bolsas de estudo.

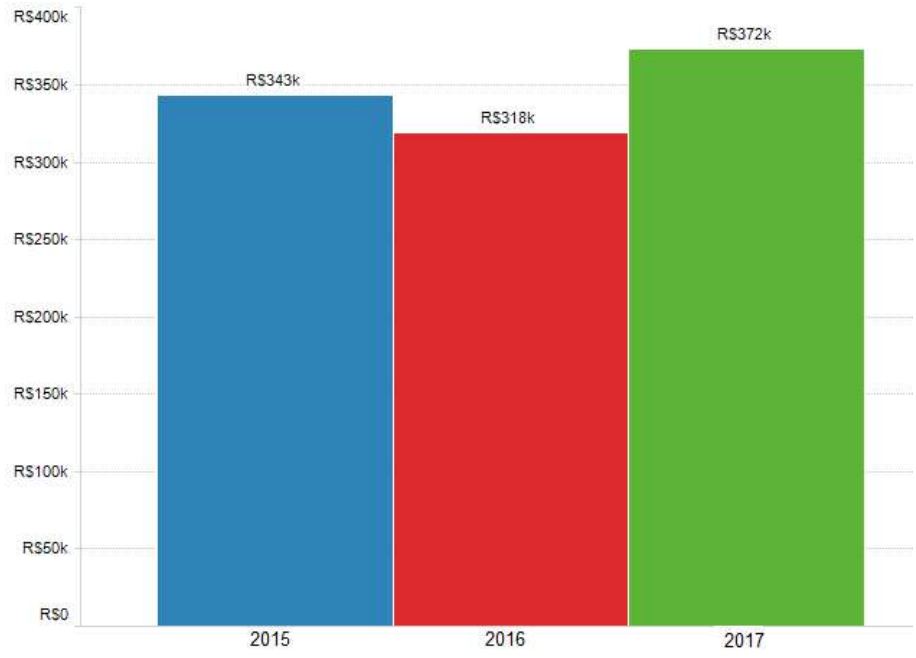

Figura 16. Orçamento para bolsa de estudo concedidas pelo campus Hortolândia nos anos 2015, 2016 e 2017.

7. Conclusão

O objetivo deste trabalho foi aplicar visualização de informação ao Portal da Transparência do IFSP tendo como base o modelo de referência de [Card et al. 1999]. Utilizou-se os conhecimentos adquiridos no curso para desenvolvimento e testes de um *web-crawler* para a coleta dos dados do Portal de Dados Abertos do IFSP. Parte dos conhecimentos foram adquiridos durante a iniciação científica que resultou no trabalho de [Oliveira and Guedes 2015].

As visualizações criadas neste trabalho demonstram que a aplicação de visualização de informação ajuda a responder perguntas e disseminar informações de modo mais eficiente. Elas podem ser utilizadas por gestores do serviço público para ajudar na tomada de decisões e criação de estratégias, pois facilitam a compreensão das informações relacionadas ao orçamento e gestão de pessoas nos *campus*. O público geral também se beneficia pois a informação é transmitida de modo mais eficaz melhorando a transparência, pois traduz os dados para uma forma mais compreensível, consequentemente atinge um público maior e estimula a população a conhecer mais sobre a gestão do IFSP.

Uma sugestão de trabalho futuro é o uso de visualizações interativas com foco no público geral.

Referências

BRASIL (2009). Lei Complementar nº 131, de 27 de Maio de 2009. `http://www.planalto.gov.br/ccivil_03/LEIS/LCP/Lcp131.htm`. [Online; acesso em 30-Agosto-2018].

BRASIL (2018). Portal da Transparência. `http://portaldatransparencia.gov.br/`. [Online; acesso em 30-Agosto-2018].

Card, S. K., Mackinlay, J. D., and Shneiderman, B., editors (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Gephi Consortium (2017). Gephi About. `https://gephi.org/about/`. [Online; acesso em 15-Setembro-2018].

Hitachi Vantara (2018). Pentaho Data Integration . `https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-data-integration.html`. [Online; acesso em 17-Setembro-2018].

IBM (2018a). IBM InfoSphere DataStage. `https://www.ibm.com/us-en/marketplace/datastage`. [Online; acesso em 17-Setembro-2018].

IBM (2018b). What is Watson Analytics? `https://www.ibm.com/watson-analytics`. [Online; acesso em 17-Setembro-2018].

Nascimento, H. A. D. and Ferreira, C. B. R. (2005). Visualização de informações – uma abordagem prática. In *Congresso da Sociedade Brasileira de Computação*, pages 1262–1312, São Leopoldo. CSBC.

Oliveira, F. S. and Guedes, G. B. (2015). Visualização de informação aplicada à legislação penal brasileira. In *VI Congresso de Iniciação Científica e Tecnológica do IFSP*, Itapetininga. IFSP.

Papaloi, A. and Gouscos, D. (2013). Parliamentary information visualization as a means for legislative transparency and citizen empowerment? *eJournal of eDemocracy and Open Government*, 5(2):174–186.

Paula, M. M. V., Ribeiro, F. C., Chaves, M., Rodrigues, S. A., and Souza, J. M. (2011). A visualização de informação e a transparência de dados públicos. In *VII Simpósio Brasileiro de Sistemas de Informação*, pages 384–395, Salvador. SBSI.

Silva, C. G. d. (2006). Exploração de bases de dados de ambientes de educação a distância por meio de ferramentas de consulta apoiadas por visualização de informação. Master's thesis, Universidade de Campinas.

TIBCO Software Inc (2018). TIBCO Spotfire Overview. `https://spotfire.tibco.com/overview`. [Online; acesso em 17-Setembro-2018].