Uso de software livre para disseminação e análise de dados abertos governamentais

Use of open source software for dissemination and analysis of open government data

Lucas Rodrigues Costa, Lucas Ângelo Silveira, Ronnie Fagundes Brito e Milton Shintaku
Instituto Brasileiro de Informação em Ciência e Tecnologia - Ibict, lucasrodrigues, lucasangelo, ronniebrito, shintaku {@ibict.br}

Resumo:
O estudo apresenta um modelo voltado à disseminação e análise de dados de órgãos governamentais e desenvolvido com software livre. Devido a grande oferta de ferramentas foi adotada uma metodologia voltada a seleção de tecnologias robustas que atendessem às demandas destas agências. Desta forma, o modelo baseou-se nos softwares *Comprehensive Knowledge Archive Network* (CKAN) e Pentaho, os quais possibilitam serviços de depósito, recuperação, visualização e análise dos dados. O caso de estudo foi aplicado na Secretaria Nacional da Juventude, onde dados oriundos de várias fontes foram recolhidos, tratados e publicados para o público em geral. O modelo atende aos objetos dos órgãos governamentais, como a oferta de acesso a dados brutos e sua análise pelo público em geral, contribuindo assim com a transparência de seus dados.
Palavras-chave: Softwares livres; dados abertos; *data warehouse*.

Abstract:
This paper presents a model for dissemination and analysis of government agencies data and which is developed with free software. Due to the great offer of tools, it followed a methodology focused on the selection of robust technologies that met agencies requirements. Thus, it resulted in a model based on the Comprehensive Knowledge Archive Network (CKAN) and Pentaho software, which offer deposit, retrieval, visualization and data analysis services. A case study was developed at the National Youth Secretariat, where data from various sources were collected, processed and published for the general public. The model attends government agencies, offering access to raw data and its analysis by the general public, contributing to transparency of the government data.
Keywords: Open source; Open data; data warehouse.

1. Introdução

No Brasil a transparência do estado tem sido promovida com várias ações, entre os quais situa-se a iniciativa para dados abertos governamentais, processo pelo qual os governantes disponibilizam informações aos seus cidadãos (OLIVERIO, 2011). Essa orientação governamental é regida pela Lei nº 12.527, que regula o acesso livre à informação governamentais, reservadas às questões de segurança e proteção dos dados sensíveis. Essa lei engloba desde dados brutos à documentos completos, tratando questões como sigilo, autenticidade, integridade e primariedade, entre outros pontos (BRASIL, 2011).

Em uma análise detalhada, Dados governamentais são descritos como resultados de atividades dos órgãos públicos e podem estar contidos em bases de dados, documentos impressos ou digitais, entre outros (OLIVEIRA, 2016), figurando assim uma grande variedade de formas e formatos. Por sua vez, dados abertos podem ser lidos, utilizados e disseminados de forma livre, sendo requisitos a citação da fonte e se for o caso o compartilhando sob mesmo tipo de licença (ISOTANI; BITTERNCOURT, 2015). Assim, dados abertos governamentais são dados abertos gerados pelo governo e que devem ser disseminados com a sociedade.

Palazzi e Tygel (2014) afirmam que parte desses dados gerados no governo são de cunho estatístico, possibilitando a geração de indicadores, que podem ser utilizados para análises e tomada de decisão. Dessa forma, para a disseminação e análise dos dados governamentais, pode-se utilizar sistemas informatizados que possibilitem a implementação de políticas relacionadas iniciativa de dados abertos governamentais para a disseminação e análise dos dados armazenados e produzidos do governo.

Com isso, tem-se oportunidades e desafios relacionados à pesquisas que atendam às necessidades das instituições e

órgãos às suas especificidades ao tratamento dos dados abertos governamentais.

Neste trabalho, apresenta-se o resultado de pesquisa efetuado na Secretaria Nacional de Juventude voltado à criação de um modelo de dados abertos governamentais utilizando software livre. Isso, contribui com a discussão sobre dados abertos governamentais e a disseminação de informação por meio de software livre.

2. Metodologia

O presente estudo tem aspectos aplicados, com utilização de técnicas ligadas à ciência da computação, na seleção e uso de tecnologias, para criação e aplicação de um modelo voltado a dados abertos com o uso de software livre. Assim, utiliza-se conceitos e técnicas voltados à avaliação de ferramentas informatizadas, alinhado à técnicas ligadas a qualidade de software. Nesse sentido, qualidade é entendida na forma da Norma ISO 8402, na qual se refere ao conjunto de características relacionadas ao atendimento das necessidades dos usuários, sejam explícitas ou implícitas. Atendendo os requisitos registrados ou não, a qualidade do software é uma avaliação de aspectos quantitativos e qualitativos, no qual contempla o processo e produto, como advogado por Tsukumo et al (1997), na medida em que o processo oferta certas garantias ao produto.

Para Silva (2007), a avaliação de softwares livres não deve ser embasada apenas na gratuidade da ferramenta, mas nos benefícios gerais que a nova tecnologia pode trazer. Seguindo tal raciocínio, o critério básico de seleção de ferramentas a serem utilizadas, era ser software livre e que provesse funcionalidades tais como: suporte por comunidade internacional atuante, mantida por instituição confiável, e que fornecesse certas garantias de sustentabilidade.

3. Ferramentas utilizadas

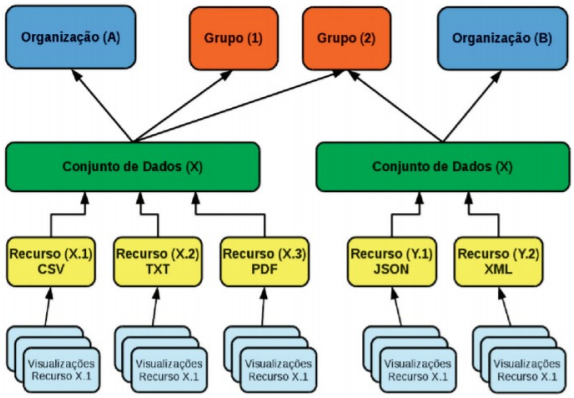
A escolha adequada das ferramentas utilizadas para a disseminação e análise dos dados do governo tem o intuito de auxiliar a iniciativa de dados abertos governamentais com o propósito de divulgar dados e informação.

Seguindo as linhas de software livres e os benefícios fornecidos chegou-se ao uso do sistema de repositórios *Comprehensive Knowledge Archive Network* - CKAN e o Pentaho. Nas próximas seções, uma descrição detalhada de ambos os softwares será realizada.

3.1. CKAN

O CKAN (CKAN, 2017) possibilita o depósito de bases de dados de forma organizada. Conforme a hierarquia descrita por Costa et al (2017) na figura 1, tem-se os conjuntos de dados como ponto central ligados às organizações. O conjunto de dados é caracterizado e descrito por metadados, podendo conter um ou mais recursos, em formato diversos. Com isso, caso tenha-se uma base de dados com arquivos em formato textuais, planilhas entre outros. Estes dados podem ser agrupados em uma única base de dados ligados a uma organização, tendo em vista que um único CKAN pode ser utilizado por uma ou mais instituições. Além disso, pode-se criar organizações artificiais para agrupar uma ou mais bases de dados denotadas como grupos e com isso facilitar a recuperação dos dados.

Figura 1: Possíveis hierarquias de bases de dados no CKAN



Este modelo alinha-se às indicações de Correa et al. (2017) as quais descrevem o uso do CKAN como uma ferramenta que apoia a disseminação de dados

governamentais, contribuindo em parte com a transparência do governo. Cabe ressaltar que o CKAN está de acordo com as orientações dos dados abertos governamentais e a Lei nº 12.527, que regula o acesso livre à informação de governo.

3.2. Pentaho

No que se refere a análise dos dados, optou-se pela ferramenta Pentaho, visto que é uma ferramenta robusta para *Business Intelligence* (BI) (AHISHAKIYE, 2017) com foco no tratamento e análise de dados (COSTA et al., 2017; MARINHEIRO; BERNARDINO, 2015).

O Pentaho oferece ferramentas de *Online Analytical Processing* (OLAP) que permitem analisar grandes volumes de dados de forma online e sob diferentes cruzamentos e dimensões dos dados, bem como a realização de cálculos complexos. O principal artefato envolvido em uma análise por meio de OLAP é o cubo multidimensional. Vale ressaltar, que a ferramenta é uma solução composta por vários módulos a fim de atender diferentes etapas de estruturação de uma base analítica. Entre os módulos disponíveis no Pentaho está o *Pentaho Data Integration* (PDI), o *Pentaho Schema Workbench* (PSW- MONDRIAN) e o SAIKU.

- **PDI -** tem por finalidade a integração de dados, possibilitando cruzar dados proveniente de várias fontes, com o uso de técnicas de ETL (extract-transform-load). O PDI oferece uma interface gráfica para a conexão das várias fontes e possibilita apresentar os resultados do processamento em formas de grafos.
- **PSW - MONDRIAN -** é uma ferramenta para o desenvolvimento de um esquema xml que descreve o cubo multidimensional dos dados.
- **SAIKU -** é um módulo para visualização dos dados do cubo de uma forma amigável e dinâmica.

4. Resultados e Discussões

A Secretaria Nacional da juventude (SNJ) possui um fluxo de dados no qual coleta ou gera uma grande quantidade de dados brutos sobre juventude, nos mais diversos temas. Conforme descreve Cury (2007), essa secretaria nasceu de uma ação interministerial, devido ao seu caráter multifacetado. O mesmo autor, relata que o tema juventude é novo na política no mundo, revelando a inovação desta secretaria e por consequência suas ações.

Após o levantamento de requisitos, verificou-se que a principal necessidade da SNJ era de um modelo que contemplasse a gestão de bases de dados estatísticas, com o fluxo informacional de Coleta ou Geração; Catalogação; e finalmente sua Recuperação/Análise. Os dados em sua maioria, são provenientes de outras instituições como o Instituto Brasileiro de Geografia e Estatística (IBGE) e Instituto de Pesquisa Estatística Aplicada (IPEA), ou gerado para o SNJ, por instituições como a Caixa Econômica Federal. Assim no primeiro caso considera-se como coleta e no segundo uma geração. Requerendo, dessa forma, uma ferramenta que possibilite o armazenamento organizado das bases de dados, permitindo a recuperação e análise dos dados.

A prospecção focou em duas etapas: a primeira consiste na migração dos dados de seus diferentes formatos para uma plataforma comum, e uma segunda que trata de disponibilizar uma ferramenta de visualização de dados e elaboração de relatórios. Pela prospecção de tecnologias foi utilizado o uso integrado do CKAN e do Pentaho.

Com isso, tem-se o CKAN como um repositório de dados, com todas as funcionalidades voltadas à gestão de bases de dados e o Pentaho como uma ferramenta com foco em analisar, organizar e apresentar tais dados.

4.1. CKAN na SNJ

Na SNJ as organizações artificiais (grupos) do CKAN tornaram-se temas de interesse da secretaria, como: saúde, educação, lazer, onde as organizações são organizadas como sendo diferentes fontes para base de dados. Com isso, tem-se um modelo de repositório para dados governamentais, organizado de forma temática, com três níveis, sendo: Tema → base de dados → recursos, possibilitando

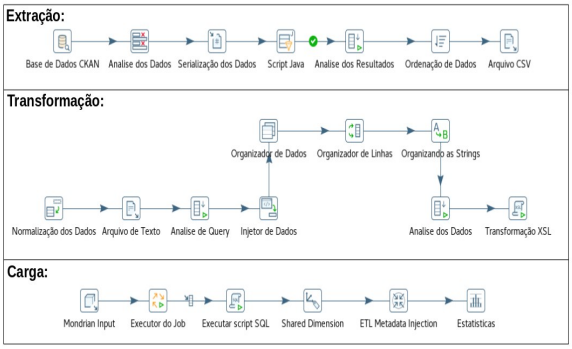
catalogar todas as bases de dados de forma organizada, facilitando a recuperação.

O CKAN neste modelo apresenta-se como repositório de dados tornando-se útil para a catalogação das bases de dados em uma estrutura organizada e integrável com outros sistemas CKAN na forma de um ecossistema de dados governamentais. Todavia, por ter um único órgão depositante de dados, tornou-se um repositório de dados institucional e temático.

4.2. Pentaho na SNJ

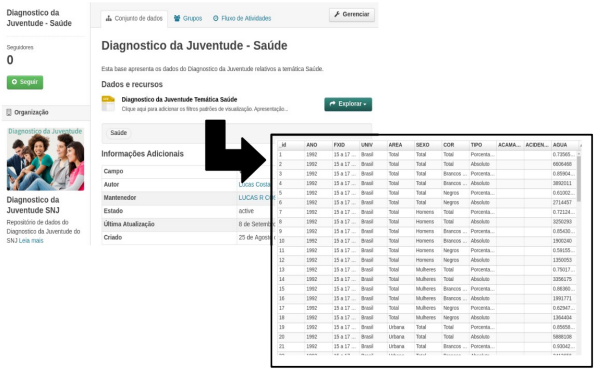
Inicialmente foi desenvolvido uma estratégia para a extração dos dados armazenados no CKAN e uma transformação dos mesmos para um formato comum, tendo em vista que diferentes bases e tipos de dados foram armazenados no repositório. Para essa tarefa foi utilizado o módulo PDI que possibilitou cruzar dados proveniente de diversas fontes do CKAN por meio das técnicas de ETL. A interface gráfica para a conexão de tais fontes auxiliou na apresentação das tarefas resultando nos processos descritos na figura 2.

Figura 2: ETL utilizado no processo da SNJ.



Como pode ser visto, foram separados as três etapas do processo ETL realizadas no PDI. A extração começa na base de dados do CKAN a qual possuem inúmeras tabelas com informações de diversas áreas relacionadas à juventude. A figura 3 mostra um exemplo do conjunto de dados relacionado a saúde da juventude no Brasil e como se encontra o formato dos dados.

Figura 3: Exemplo de conjunto de dados da SNJ.

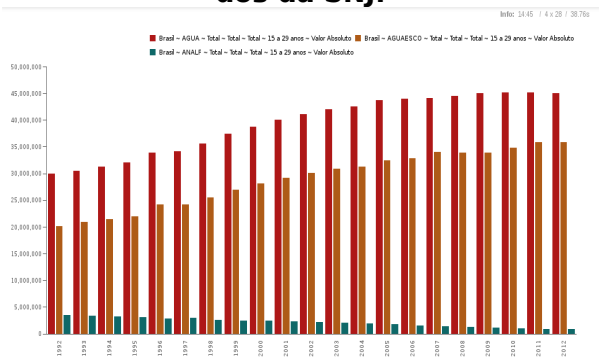


Ainda na figura 2, após a extração, tem-se o processo de transformação dos dados. Nesta etapa todas os dados são convertidos para um formato e estrutura comum. Em seguida, tem-se a parte da carga em que realiza-se a inserção de dados em um banco de dados relacional com uma estrutura específica para a criação do cubo.

Posteriormente, o PSW-MONDRIAN foi utilizado para a criação do cubo, o qual permite análise multidimensional de dados podendo-se desenvolver diferentes soluções para exploração do repositório de dados por meio de uma ferramenta de análise para a produção de informações em diversos formatos, tais como, tabela, gráficos e relatórios.

Por fim, utiliza-se o SAIKU para a visualização dos dados conforme a orientação do cubo possibilitando a busca de informação de forma dinâmica. A figura 4 mostra um exemplo em que foram relacionados três bases de dados do CKAN, relacionando jovens com acesso adequado a água, saneamento e sua relação com a taxa de analfabetismo. A informação extraída deste gráfico mostra que jovens com acesso a água e saneamento apresentam taxa de analfabetismo menor.

Figura 4: Exemplo de conjunto de dados da SNJ.



Vale ressaltar que as consultas podem ser realizadas de forma dinâmica cruzando vários tipos de dados em diversos tipos de gráficos, relatórios e planilhas. Dessa forma, com o SAIKU, pode-se reutilizar os dados do CKAN para geração de novas informações, incrementando as possibilidades de uso das bases de dados mantidas no repositório.

O modelo apresentado no presente estudo encontra semelhanças na experiência de Tygel (2012), na medida em que utiliza o Pentaho para análise de dados abertos governamentais. De Faria Cordeiro et al (2011) utilizaram o Pentaho para integrar uma base de dados para uso de ferramentas semânticas. Mendonça, Cruz e Campos (2014) utilizaram o Pentaho como parte de um modelo para tratamento de dados governamentais. Dos Santos e da Silva (2014) comparam o sistema I-GOV com o uso do Pentaho para integrar dados governamentais. Portanto, o Pentaho revela-se apropriado ao apoio à análise de dados governamentais de características estatísticas.

O resultado do caso de estudo pode ser encontrado em: http://magonia.ibict.br/ckan/consultas-livres.

5. Considerações Finais

O artigo apresenta um estudo de caso com os dados da Secretaria Nacional da Juventude (SNJ) para a colaboração da proposta.

A contribuição do modelo utilizado no presente estudo está no uso de duas ferramentas livres (CKAN e o Pentaho) para atender as necessidades da SNJ para com às orientações dos dados abertos governamentais e ofertar aos gestores de políticas de juventude bem como a sociedade em geral, um sistema web com oferta de dados brutos ou pré-analisados. O modelo desenvolvido e implementado na SNJ composto pelo CKAN e Pentaho se apresentou eficaz, dinâmico podendo ser implementado em outros órgãos de governo, contribuindo com a discussão do uso de ferramentas livres em órgãos públicos. Além de se apresentar eficaz no atendimento aos objetivos do estudo. O CKAN possibilita a criação de um repositório, com a catalogação de bases de dados de forma organizada. Para facilitar a recuperação o Pentaho possibilita cruzar dados das bases armazenadas no CKAN para a integração e elaboração de novas informações.

Apresentou recursos tecnológicos que apoiam a gestão estratégica e eficiente das organizações. Para isso é proposto uma metodologia de integração baseada nos softwares livres CKAN e Pentaho para ajuda na tomada de decisões organizacionais.

A metodologia dos dois sistemas combinados é capaz de fortalecer o plano de atuação das organizações através da geração de informações rápidas, precisas e personalizáveis garantindo uma estruturação de gestão diferenciada para a melhora no processo de tomadas de decisões pelos gestores organizacionais.

6. Referências

ALBANO, Cláudio Sonaglio. Dados governamentais abertos: proposta de um modelo de produção e utilização de informações sob a ótica conceitual da cadeia de valor. 2014. Tese de Doutorado. Universidade de São Paulo.

AHISHAKIYE, Emmanuel et al. Comparative Analysis of Open source Business Intelligence tools for Crime Data Analytics.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011 .

CKAN. Documentation. 2013. Disponível em: <http://docs.ckan.org/en/latest/>. Acesso em: 18 jul. 2017.

COSTA, Lucas Rodrigues et al. Guia do usuário CKAN. IBICT. 2017.

COSTA, Evandro B. et al. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, v. 73, p. 247-256, 2017.

CORRÊA, Andreiwid Sh et al. Transparency and open government data: a wide national assessment of data openness in Brazilian local governments. Transforming Government: People, Process and Policy, v. 11, n. 1, p. 58-78, 2017.

CURY, BETO. Admirável mundo novo. In: KEHL, Maria Rita. A juventude como sintoma da cultura. Revista de saberesmandato vereador Arnaldo Godoy, p. 44-55, 2007.

DE FARIA CORDEIRO, Kelli et al. An approach for managing and semantically enriching the publication of Linked Open Governmental Data. In: Proceedings of the 3rd workshop in applied computing for electronic government (WCGE), SBBD. 2011. p. 82-95.

DE OLIVEIRA, Carolina. A gestão arquivística de documentos como apoio à publicação de dados governamentais abertos. Acervo, v. 29, n. 2 jul-dez, p. 168-178, 2016.

DA SILVA, José Fernando Modesto; SUBSÍDIO À GESTÃO BIBLIOTECÁRIA. CBBD.

DOS SANTOS, João Paulo Clarindo; DA SILVA, Fábio José Coutinho. IGOV: um sistema de integração de dados governamentais. Revista Brasileira de Administração Científica, v. 5, n. 2, p. 8-16, 2014.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. Dados Abertos Conectados: Em busca da Web do Conhecimento. Novatec Editora, 2015.

MARINHEIRO, Antonio; BERNARDINO, Jorge. Experimental evaluation of open source business intelligence suites using OpenBRR. IEEE Latin America Transactions, v. 13, n. 3, p. 810-817, 2015.

DE MENDONÇA, Rogers Reiche; CRUZ, S. M. S.; CAMPOS, Maria Luiza M. Gerência de proveniência multigranular em linked data com a abordagem etl4linkedprov. Anais do Simpósio Brasileiro de Bancos de Dados. Paraná, Brazil, 2014.

OLIVERIO, Marcio Araujo. Governo aberto como ferramenta de comunicação entre o governo e o cidadão. In: XXXIV Congresso Brasileiro de Ciências da Comunicação. Recife, PE. 2011.

PALAZZI, Daniele; TYGEL, Alan. Visualização de Dados Estatísticos Representados como Dados Abertos Ligados. 2014.

TSUKUMO, Alfredo N. et al. Qualidade de software: visões de produto e processo de software. II ERI–SBC, Piracicaba, São Paulo, Brasil, 1997.

TYGEL, Alan. Representação e Visualização de dados estatísticos: os desafios dos dados abertos ligados. 2012