Geração Semiautomática de Gráficos para Jornalismo de Dados Usando Dados Abertos: Um Estudo de Caso do Censo da Educação Superior

Semiautomatic Generation of Graphics for Data Journalism Using Open Data: A Case Study of Higher Education Census

Felipe C. P. Magalhães
Licenciado em Computação, Trabalho de Conclusão de Curso - Departamento de Ciência da Computação, Universidade de Brasília. Trabalho de Final de Curso com orientação Prof. Dr. Edison Ishikawa e coorientação Profa. Dra. Suzana Guedes Cardoso. E-mail: fcpmagalhaes@hotmail.com

Prof. Dr. Edison Ishikawa
Doutor em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro e Mestre em Informática pela Pontifícia Universidade Católica do Rio de Janeiro. Engenheiro de Computação pelo Instituto Militar de Engenharia. Professor Adjunto do Departamento de Ciência da Computação da Universidade de Brasília. E-mail: ishikawa@unb.br

Profa. Dra. Suzana Guedes Cardoso
Pós- Doutora. Professora Associada do Curso de Jornalismo, Departamento de Jornalismo, Faculdade de Comunicação, Universidade de Brasília. E-mail: suzanagc@gmail.com

Resumo

O Jornalismo Computacional refere-se às ferramentas e algoritmos que os jornalistas usam para contar histórias. Neste contexto, os dados abertos podem servir como fonte da notícia. Embora diferentes técnicas de manipulação de dados, como Data Warehouse, Data Mining e Business Intelligence, possam responder a perguntas sobre dados, suas respostas geralmente são estáticas e não interativas, limitando-se a questões predefinidas. Este projeto teve como objetivo desenvolver uma aplicação que permitisse a geração de gráficos dinâmicos a partir da seleção instantânea de filtros aplicados aos dados do Censo da Educação Superior do Brasil. Essa abordagem tinha o propósito de facilitar a interpretação desses dados por jornalistas. O protótipo implementado foi avaliado usando a escala SUS e obteve resultados satisfatórios em relação à usabilidade do sistema. Como resultado, essa abordagem possibilitou a extração de novas informações que podem ser utilizadas na criação de matérias jornalísticas.

Palavras-chave: Jornalismo de Dados, Dados Abertos, Censo da Educação Superior, INEP, Gráficos.

Abstract

Computational Journalism refers to the tools and algorithms that journalists use to tell stories. In this context, open data can serve as a news source. Although various data manipulation techniques, such as Data Warehouse, Data Mining, and Business Intelligence, can answer questions about data, their responses are often static and non-interactive, limited to predefined queries. This project aimed to develop an application that allows the generation of dynamic graphs from the instant selection of filters applied to data from the Higher Education Census of Brazil. This approach was intended to facilitate the interpretation of this data by journalists. The implemented prototype was evaluated using the SUS scale and achieved satisfactory results in terms of system usability. As a result, this approach enabled the extraction of new information that can be used in the creation of news articles.

Keywords: Data Journalism, Open Data, Higher Education Census, INEP, Graphics.

Artigo recebido em: 12/07/2023 e Aprovado em: 01/11/2023

1. INTRODUÇÃO

As novas formas de armazenamento de dados na web têm permitido a descoberta e a publicação de informações que possibilitam a geração de novos conhecimentos acessíveis a toda a população. Isso antes era restrito por se tratar de informações limitadas a determinados grupos de pessoas (Carneiro, 2016). Essas informações, quando disponibilizadas de forma livre, são classificadas como conhecimento aberto. Segundo a Open Definition (Doctorow *et al*., 2018), o conhecimento é considerado aberto se qualquer pessoa está livre para acessá-lo, utilizá-lo, modificá-lo, e compartilhá-lo – restrito, no máximo, a medidas que preservam a proveniência e a abertura.

Esse tipo de conhecimento aberto tem crescido exponencialmente à medida que a humanidade está cada vez mais conectada às redes de intercomunicação como a internet, produzindo e consumindo informações numa velocidade nunca antes vista. A International Data Corporation – IDC (Reinsel *et al*., 2017) prevê que, até o ano de 2025, a esfera de dados global aumentará para 163 zettabytes, o que corresponde a mais de um trilhão de gigabytes. Seguindo a mesma ótica de conhecimento aberto, os dados abertos, para serem assim classificados, precisam ser livremente utilizados, reutilizados e redistribuídos, sendo necessário apenas citar a fonte (Dietrich *et al*., 2018).

Com base nesse contexto, este trabalho visou explorar os dados abertos do Censo da Educação Superior do Brasil, fornecidos anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP. Esses dados incluem informações sobre as instituições de ensino superior do país, seus cursos, docentes e alunos. O propósito desses indicadores, segundo o próprio INEP, é fornecer à comunidade acadêmica e à sociedade em geral informações detalhadas sobre a situação e as principais tendências do setor educacional.

A análise desses dados viabiliza a geração de novas conhecimentos por meio de consultas voltadas para o campo do jornalismo computacional. Ao cruzar dados selecionados pelos usuários, são gerados gráficos com o intuito de tornar a compreensão desses números mais acessível, apresentando informações visuais por meio de modelos gráficos. O objetivo principal é ampliar o entendimento estatístico obtido a partir do censo.

A vasta quantidade de informações gerada atualmente por meio de sistemas computacionais difere da produção humana, em que os dados coletados frequentemente estão diretamente ligados aos eventos e suas fontes. Muitas vezes, essas informações se reduzem a

dados numéricos representados por sequências de 0 e 1, que são processadas automaticamente e, em muitos casos, podem ser incompreensíveis para o entendimento comum.

As informações qualitativas proporcionam aos pesquisadores a oportunidade de conduzir análises aprofundadas e específicas, visto que envolvem contato direto com os sujeitos da pesquisa, que são entrevistados e observados em suas ações. Mediante as opiniões e perspectivas desses indivíduos, os dados qualitativos são coletados. Em contrapartida, os dados abordados neste estudo são de natureza quantitativa. Isso implica que esses dados são principalmente estatísticos, oferecendo um maior potencial para extrair inferências, identificar padrões e tendências. Ao mesmo tempo que são mais difíceis de serem compreendidos tanto por seu pesquisador quanto pelo público-alvo, são também criticados por reduzirem seu conteúdo a modelos extremamente simplistas (Carneiro, 2016).

Desta forma, tornou-se necessário criar uma interface amigável para os jornalistas, com o objetivo de facilitar a transformação desses números em informação visual de fácil compreensão. Isso não exige que o usuário tenha conhecimentos avançados em informática.

2 DADOS ABERTOS GOVERNAMENTAIS

Conforme a definição da Open Knowledge Foundation (Doctorow *et al*., 2018), “dados abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras”. Os dados disponibilizados pelo INEP, que são objetos deste estudo, classificam-se como dados abertos governamentais.

Segundo a World Wide Web Consortium – W3C (W3C, 2018), “Dados Abertos Governamentais são a publicação e a disseminação das informações do setor público na web, compartilhadas em formato bruto aberto, compreensíveis logicamente, de modo a permitir sua reutilização em aplicações digitais desenvolvidas pela sociedade”.

Inspirados pela filosofia do código aberto, fundamentada nos princípios de abertura, participação e colaboração, compreendemos que os dados governamentais são propriedades comuns, assim como as informações científicas.

3 JORNALISMO DE DADOS

O jornalismo de dados desempenha um papel fundamental ao auxiliar jornalistas na narrativa de histórias complexas por meio de gráficos que prendam a atenção. Os dados podem servir como fonte primária para o jornalismo de dados ou como ferramentas para contar histórias (Gray *et al*., 2012). Desta forma, pode ser definido como um conjunto de práticas para coletar, analisar, visualizar e publicar dados para fins jornalísticos, conforme definem Berret e Philips (2016).

Outra definição de Howard coloca o jornalismo lado a lado com dados científicos e o define como uma "aplicação da ciência de dados ao jornalismo, em que a ciência de dados é definida como o estudo da extração de conhecimento a partir de dados" (Howard, 2014). Howard considera o jornalismo de dados como um processo que engloba reunir, limpar, organizar, analisar, visualizar e publicar dados para apoiar a criação de conteúdo jornalístico. Portanto, o jornalismo de dados promove uma abordagem científica e baseada em fatos. Essa abordagem exige que o jornalismo seja tratado e praticado da mesma forma que as investigações científicas, convidando a métodos científicos, objetividade científica, transparência científica e reprodutibilidade científica. Esses métodos científicos incluem análise de dados quantitativos e qualitativos para investigação, produção de conteúdo jornalístico e comunicação desses conteúdos ao público.

4 MODELAGEM

A estrutura escolhida para a implementação do protótipo deste trabalho é também a mais difundida na arquitetura multicamadas, conhecida como *three-tier*. Nesse padrão de arquitetura, há subsistemas separados para as seguintes camadas (Fowler, 2002):

- Cliente: Camada responsável pela exibição da interface do usuário. Por meio dela, o usuário é capaz de interagir com a aplicação.
- Lógica de Negócios: Camada responsável pelo controle de toda a lógica de negócios, bem como da comunicação com o banco de dados.
- Dados: Camada responsável pelo armazenamento de todos os dados necessários para o funcionamento da aplicação.

5 MANIPULAÇÃO DOS DADOS

Os microdados do Censo da Educação Superior reúnem o conjunto de informações detalhadas sobre o censo, que são disponibilizados anualmente no portal do INEP e organizados em cinco módulos distintos: Aluno, Curso, IES (Instituição de Ensino Superior), Local de Oferta e Docente. Cada módulo representa o conjunto de dados coletados por meio de questionários correspondentes a sua entidade equivalente, bem como métricas armazenadas em variáveis derivadas que resultam do processo inicial de coleta de dados.

Como primeira etapa para o desenho do mapa conceitual, foi realizada a análise campo a campo de cada módulo, a fim de estabelecer correlação entre eles. O mapa conceitual elaborado (Magalhães *et al*., 2023) foi gerado a fim de apresentar uma visão horizontal do problema e trazer clareza quanto à complexidade dos dados contidos nos microdados.

Identificou-se que a entidade Aluno possui o maior número de atributos de interesse a serem consultados e extraídas novas informações. Com isso, as entidades Local de Oferta e Docente foram desprezadas para este estudo, pois não possuíam ligação direta com o objeto central de pesquisa.

O grande volume de dados coletados em todo o território nacional tornou o tamanho final de cada uma das planilhas na casa dos megabytes, o que já são tamanhos consideráveis para o formato ".csv". Em particular, as tabelas relacionadas à entidade Aluno, que é o foco principal da análise de informações, alcançaram tamanhos da ordem de gigabytes.

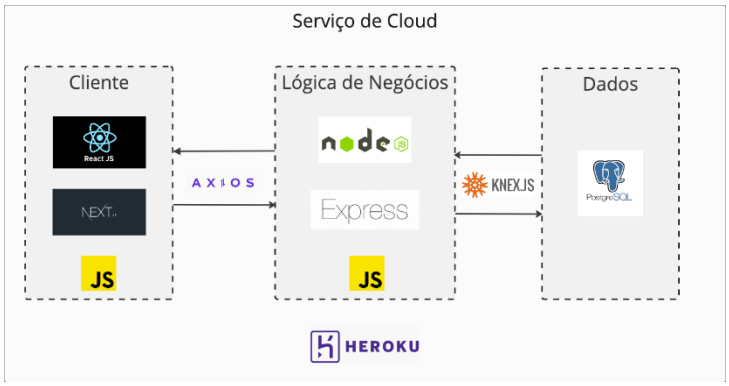
A manipulação de tabelas com volumes tão grandes, como as mencionadas, torna-se inviável em computadores de uso pessoal, como os utilizados pelos orientadores e orientando deste trabalho. Portanto, a estratégia adotada foi a fragmentação desses arquivos originais em arquivos menores, com um limite de 10 MB, antes de serem carregados na base de dados física. Um exemplo é a tabela de cursos do ano de 2019, que contém um total de 40.709 registros, referentes aos cursos vinculados às Instituições de Ensino Superior cadastradas junto ao MEC. Já a tabela de Alunos do mesmo ano, a mais volumosa em número de registros, abriga dezenas de milhões de dados, pois inclui todos os alunos vinculados a uma IES. Dado o objetivo deste trabalho de apresentar uma linha do tempo entre os anos selecionados pelo usuário, tornou-se inviável armazenar toda essa massa de dados em um servidor, devido aos custos operacionais elevados e ao tempo de latência das consultas SQL necessárias para varrer todos esses dados antes de atingir o resultado final, que é a exibição dos gráficos.

Como medida para mitigar esses problemas, foi realizada uma adequação em nível de protótipo, tomando como medida a redução do escopo dos dados consultados no projeto. Em vez de trabalhar com dados de abrangência nacional, passamos a nos concentrar exclusivamente nas informações relacionadas ao Distrito Federal. Para isso, foram aplicados filtros em todas as tabelas das entidades carregadas, a fim de armazenar registros somente de Alunos vinculados a Instituições de Ensino Superior sediadas no Distrito Federal, e Cursos vinculados a essas mesmas IESs.

6 PROTÓTIPO

As tecnologias empregadas para a construção do artefato foram selecionadas com o objetivo de possibilitar o acesso à aplicação por meio de uma interface Web. A Figura 1 exibe as camadas de arquitetura com as respectivas tecnologias escolhidas para seu desenvolvimento.

Figura 1 – Arquitetura *three-tier* e tecnologias utilizadas no protótipo



Fonte: Elaboração própria.

A Camada de Lógica de Negócios foi escrita usando Node.js e Express.js. Essa camada representa o Servidor da Aplicação que atua como a ponte de comunicação entre a Camada de Cliente e a Camada do Banco de Dados. Ela orquestra toda a dinâmica de quais filtros podem ser aplicados em cada etapa de consulta, realiza o cruzamento das informações e retorna os valores finais que são utilizados para criação de gráficos na última etapa da funcionalidade da aplicação. A biblioteca Knex.js foi utilizada para fazer a comunicação com a Camada de Banco de Dados.

A Camada de Cliente, foi escrita utilizando a biblioteca JavaScript React.js, com o framework Next.js. Essa camada representa o Cliente da Aplicação, ou seja, ela consome e exibe todas as informações advindas da Camada de Lógica de Negócios para o usuário final. Nela é possível realizar toda a interação com as páginas da web e acessar todas as funcionalidades do projeto.

7 RESULTADOS

Nos primeiros ensaios sobre este trabalho, que se originou em 2019, os microdados do CES eram disponibilizados somente por planilhas eletrônicas em formato (".csv"), cabendo aos jornalistas manipulá-las para extrair informações ou restringir-se às informações divulgadas nas coletivas de imprensa anuais para noticiar as últimas estatísticas. Nos anos que se seguiram, o INEP fez o lançamento de dois portais para divulgação dos dados, sendo o mais recente em dezembro de 2022. O Painel de BI do Censo da Educação Superior apresenta as principais estatísticas e seus indicadores resultantes, objetivando facilitar a consulta dos usuários a essas informações. Com propósitos parecidos mas aplicabilidades diferentes, foi elaborado então um modo para dimensionar a percepção de usuários ao fazerem uso do Painel de BI lançado pelo INEP e o portal protótipo desenvolvido neste trabalho.

O System Usability Scale – SUS é uma escala criada em 1986 por John Brooke, cujo objetivo é avaliar a efetividade (se o usuário consegue completar o objetivo), a eficiência (quanto esforço e recursos são necessários para isso) e a satisfação (se a experiência foi satisfatória) de um sistema ou aplicação. Para isso, utiliza-se uma escala de 1 a 5, em que 1 representa “discordo totalmente” e 5 representa “concordo totalmente”. Posteriormente, essas pontuações são calculadas para se obter um resultado final. Para o teste de usabilidade, foi aplicado um questionário aos usuários, solicitando que realizassem uma consulta em cada aplicação, a fim de avaliar a facilidade de navegação em cada plataforma. Desses avaliadores, 40% eram jornalistas, enquanto os restantes 60% eram usuários comuns com interesse pelo tema.

Os seguintes cenários foram descritos:

Caso de Uso 1 - Desejo realizar uma pesquisa contabilizando quantos alunos pretos, pardos e indígenas do sexo masculino se formaram no curso de Computação Licenciatura na

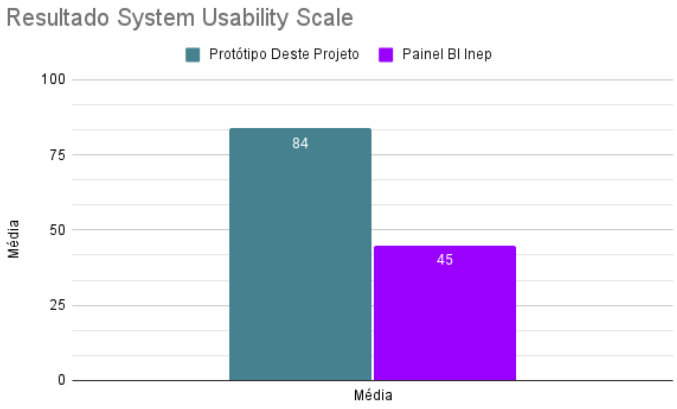
Universidade de Brasília entre os anos de 2018 e 2019, disponibilizando essas informações por meio de um gráfico de barra.

Caso de Uso 2 - Desejo realizar uma pesquisa para identificar alunos brancos e pretos/pardos que se formaram no curso de Computação Licenciatura na Universidade de Brasília entre os anos de 2018 e 2019, apresentando essas informações por meio de um gráfico de pizza com informações percentuais.

O primeiro caso de uso foi aplicado ao protótipo desenvolvido neste projeto. Já o segundo caso de uso, foi executado dentro do Painel de BI do INEP, disponível em: https://app.powerbi.com/view?r=eyJrIjoiNjUzZjU2YzItY2VlZC00MzcwLTk4OWYtODMzNWEyNzJkM2ZhIiwidCI6IjI2ZjczODk3LWM4YWMtNGIxZS05NzhmLWVhNGMwNzc0MzRiZiJ9. A avaliação dos usuários nas duas plataformas pode ser vista lado a lado em seguida.

Como resultado do cálculo realizado pela escala System Usability Scale, pôde-se observar que a média adquirida pela aplicação do protótipo foi de 84, enquanto a média adquirida pelo Painel de BI foi de 45. Esses valores não são expressos em porcentagem, mas sim em uma escala de 0 a 100, na qual a média de uma classificação positiva é de 68 pontos.

Figura 2 – Resultados Finais - Teste de Usabilidade



Fonte: Elaboração própria.

8 CONCLUSÃO

Os dados do Censo da Educação Superior fornecidos pelo INEP, assim como muitos outros dados públicos disponibilizados na web, carecem de uma estrutura dinâmica capaz de

facilitar a sua consulta, tornando as informações mais acessíveis tanto para jornalistas quanto para a população em geral, que não possui conhecimentos avançados em computação

O objetivo deste trabalho foi criar um protótipo de aplicação web de fácil compreensão para o usuário, com um fluxo intuitivo que permitisse uma navegação fluída e autodidata. Os resultados obtidos por meio da aplicação do teste de usabilidade SUS comprovaram que a abordagem adotada para resolver esse problema, por meio da construção de um projeto dedicado utilizando uma arquitetura em três camadas, atendeu ao objetivo inicial deste trabalho.

9 REFERÊNCIAS

BERRET, Charles; PHILLIPS, Cheryl. (2016). *Teaching data and computational journalism*. Columbia Journalism. Disponível em: https://www.gitbook.com/book/columbiajournalism/teaching-data-computational-journalism/detail. Acesso em: 09 out. 2023.

CARNEIRO, Márcio. *Comunicação digital e jornalismo de inserção*: como big data, inteligência artificial, realidade aumentada e internet das coisas estão mudando a produção de conteúdo informativo. São Luís: Labcom Digital, 2016.

DIETRICH, Daniel; GRAY, Jonathan; McNAMARA, Tim; POIKOLA, Antti; POLLOCK, P.; TAIT, Julian; ZIJLSTRA, Ton. *Open Data Handbook*. (s.d.). Disponível em: http://opendatahandbook.org/guide/pt_BR/what-is-open-data/#o-que-é-aberto. Acesso em: 09 out. 2023.

DOCTOROW, Cory; SUBER, Peter; HUBBARD, Tim; MURRAY-RUST, Peter; WALSH, Jo; TSIAVOS, Prodromos; MOELLER, Erik. *Open Definition*. (s.d.). Disponível em: https://opendefinition.org/od/2.0/pt-br/. Acesso em: 09 out. 2023.

FOWLER, Martin. *Patterns of Enterprise Application Architecture*. Boston: Addison Wesley, 2002.

GRAY, Jonathan; CHAMBERS, Lucy; BOUNEGRU, Liliana. *The Data Journalism Handbook*: How journalists can use data to improve the news. 2012. Newton: O'Reilly Media, 2012.

HOWARD, Alexander B. *The art and science of data-driven journalism*. Columbia Journalism School. 2014. Disponível em: http://towcenter.org/wp-content/uploads/2014/05/Tow-Center-Data-Driven-Journalism.pdf. Acesso em: 09 out. 2023.

MAGALHÃES, Felipe C. P.; CARDOSO, Suzana G.; ISHIKAWA, Edison. *Geração Semi Automática de Gráficos Para Jornalismo de Dados Usando Dados Abertos:* Um Estudo de Caso do Censo da Educação Superior. ApliCon23. 2023.

REINSEL, David; GANTZ, John; RYDNING, John. *Data age 2025*: The evolution of data to life-critical. IDC White Paper. 2017. Disponível em: https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf. Acesso em: 09 out. 2023.

W3C – World Wide Web Consortium. *Dados abertos governamentais*. 1994. Disponível em: http://www.w3c.br/divulgacao/pdf/dados-abertos-governamentais.pdf. Acesso em: 09 out. 2023.