

CONSTRUÇÃO DO JUSBD PARA FACILITAR O ACESSO AOS DADOS DO PODER JUDICIÁRIO

Weverton Ryan Ribeiro da Mata ¹; Michele A. Brandão ²; Ederson dos Santos Ramalho ³

1 Weverton Ryan Ribeiro da Mata, Técnico em Informática, IFMG Campus Ribeirão das Neves, Ribeirão das Neves - MG; ryanmata40@gmail.com

2 Michele A. Brandão, Pesquisadora do IFMG, Campus Ribeirão das Neves, michele.brandao@ifmg.edu.br

3 Ederson dos Santos Ramalho, Pesquisador do IFMG, Campus Ribeirão das Neves, ederson.ramalho@ifmg.edu.br

RESUMO

A forense digital é uma ciência que visa a identificação, preservação, recuperação, análise e apresentação de evidências de crimes em dados. Tal ciência também auxilia a apoiar ou refutar algum tipo de suposição. No Brasil, a Lei de Acesso à Informação (Lei nº 12.527) possibilita que dados públicos sejam auditados e até mesmo passem por uma análise forense. Entretanto, muitos desses dados estão em PDF (Portable Document Format) ou em planilhas grandes com várias abas, o que dificulta qualquer tipo de análise. Assim, este trabalho representa um avanço na análise de dados públicos ao baixar os arquivos disponíveis online, processar e tratar tais dados, unir dados de diferentes bases e modelar um banco de dados relacional para bases de dados do poder judiciário brasileiro disponibilizadas no site do Conselho Nacional de Justiça. O novo banco de dados é chamado de JusBD e permite a obtenção de diversas informações sobre o poder judiciário, tais como local de lotação dos magistrados, contracheques, possíveis indenizações, entre outras. O novo banco de dados também possibilita análises forenses, por exemplo, ao permitir investigar sobre os salários dos magistrados e até mesmo sobre o pagamento indevido de indenizações. Além disso, o JusBD também armazena a estrutura de uma rede social entre magistrados, na qual os indivíduos são conectados se trabalharam em um mesmo tribunal. Tal rede permite, por exemplo, identificar similaridades e discrepâncias entre magistrados de um mesmo tribunal e de tribunais diferentes. Ademais, essa rede social viabiliza os próximos passos desta pesquisa que é a inserção de informações extraídas de mídias sociais de forma a complementar os dados de interações já existentes entre magistrados com o objetivo de analisar o fluxo de interações entre indivíduos para definir indicadores capazes de classificar tais interações. Finalmente, o JusBD pode ser facilmente incrementado com dados de outras fontes, tais como redes sociais online, mídias sócias, websites, plataformas governamentais, entre outras.

INTRODUÇÃO:

Nas últimas décadas, um tipo de crime que tem se difundido é o ocorrido em meios eletrônicos ou digitais. Identificação não autorizada de dados de saúde de pacientes [Anderson and Williams 2018], ocorrência de fraudes bancárias [Becker et al. 2017] e roubos de identidade [Dadkhah et al. 2018] são exemplos de tais crimes. O grande volume de informações ao qual a internet expõe os indivíduos todos os dias e a popularização dos dispositivos digitais são algumas das causas para essa difusão. Se por um lado esses dois fatores contribuem para o aumento do número de vítimas de crimes, por outro lado, também aumentam a quantidade de conteúdo, atividades e rastros, que podem servir de evidências deixadas por ofensores.

Os dados sobre tais evidências podem ser identificados, preservados, recuperados, analisados e apresentados por uma ciência denominada Forense Digital [Guarino 2013]. Com frequência, essas evidências são relacionadas a crimes virtuais. Além disso, a forense digital auxilia a apoiar ou refutar algum tipo de suposição, reconstruir eventos criminais e prever ações não autorizadas [Guarino 2013]. Por exemplo, Greengard [2012] descreve o caso de um funcionário do Burger King em Ohio (estado dos Estados Unidos) que postou fotos em uma rede social com os pés em bandejas de alfaces. Diferentes usuários da rede social fizeram uma simples investigação forense e descobriram a localização do Burger King. Como resultado, três funcionários foram demitidos.

Nesse contexto, a Lei nº 12.527, também conhecida como “Lei de Acesso à Informação”, foi sancionada em 18 de novembro de 2011 com o objetivo de promover a Transparência da Administração Pública Brasileira. Tal lei foi modificada pelo Decreto nº 9.690, de 23 de janeiro de 2019, que aumentou a quantidade de pessoas aptas a decidirem sobre o sigilo de dados públicos. Apesar dessa mudança, ainda há diversas bases de dados públicos disponíveis na Web. Entretanto, muitos desses dados estão em PDF (Portable Document Format) ou em uma grande planilha, muitas vezes, com várias abas. Isso dificulta uma auditoria nesses dados e até mesmo uma análise forense. Assim, este trabalho representa um avanço para a análise de dados públicos ao baixar, processar, unir diferentes planilhas e modelar um banco de dados relacional para bases de dados públicas do poder judiciário brasileiro disponibilizadas no site do Conselho

Nacional de Justiça (CNJ)¹. Também são adicionadas informações de relacionamentos sociais entre magistrados que podem auxiliar em uma análise forense. Ademais, tais dados podem ser enriquecidos posteriormente com informações de redes e mídias sociais.

METODOLOGIA:

Inicialmente, para a construção do JusBD, houve uma verificação de dados que poderiam ser extraídos no site do CNJ. Desse modo, após as buscas na plataforma, foram selecionadas seis possíveis bases de dados com formato como CSV (Comma-separated Values), .xls ou com tabelas para a montagem do Banco de dados. Com posse das bases de dados, primeiramente se verificou ao que tais dados se referiam, chegando a uma exclusão de três possíveis bases para a criação do conjunto. Dessas três bases de dados, dois se mostraram inutilizáveis por existir uma limitação de estarem em uma plataforma denominada Qlikview², que apesar de algumas tentativas, não se conseguiu obter dados dessa plataforma de forma organizada e/ou sucinta. Já a terceira base, que não se mostrou eficiente para ser utilizada, foi um relatório do censo do poder judiciário, onde foi possível encontrar apenas uma tabela em formato .xls com a quantidade e porcentagem de respostas por tribunal. Os dados que se não foram eliminados por essa seleção, são dados referentes a remuneração dos magistrados, estatísticas feitas anualmente desde 2004 e a base de dados do primeiro e segundo balanço sócioambiental do poder judiciário.

Para o encetamento do conjuntos se utilizou o MySQL Workbench para a construção do Banco de dados, HeidiSQL (um programa livre e cliente código-aberto para o MySQL) para consulta e importação dos arquivos CSV para as tabelas e um terceiro software para conversão de formato .xls para CSV. Na primeira etapa para a construção do conjunto de dados iniciou com a criação de diversas tabelas. Para essa construção, analisando os arquivos que seriam importados, verificou-se que para colunas que o valor deveria ser numérico, haviam alguns caracteres com “R\$” que impediam tal inserção de um tipo int, float entre outros na coluna, então todas as colunas com tais problemas em todas as bases de dados tiveram que iniciar com o atributo varchar para no final da construção efetuar *replace* (substituição) de tais caracteres que não corresponde a o tipo de atributos desejados.

Após a preparação das tabelas, em uma segunda etapa, foi realizada a alteração dos dados em um formato para a importação nas tabelas que foi feita por meio da utilização de um software de conversão. Assim, foram convertidos cerca de 1.523 arquivos .xls com cinco tabelas cada gerando 7.615 arquivos .csv referentes a remuneração dos magistrados, dois arquivos .xls referente ao primeiro e segundo balanço sócioambiental do poder judiciário e por fim um arquivo .xls de estatísticas oficiais do poder judiciário.

Na terceira etapa, iniciou-se a importação desses dados, na qual houveram vários. A primeira tentativa de inserir esses dados foi feita no MySQL Workbench pela função `LOAD LOCAL INFILE` que por algumas irregularidades nos arquivos convertidos, não foi possível, assim tentando de outras formas por meio de scripts Python e Java, mas novamente sem sucesso, desta maneira acarretando na descoberta do HeidiSQL que na sua função para importação de arquivos tem um maior controle sobre as tabelas, mas com a desvantagem de ter de fazer a importação dos dados um por um, desse jeito demandando muito tempo pela quantidade de dados.

Após esses processos, foi feita uma modelagem relacional (que pode ser visualizado na Figura 1) para o banco de dados JusBD, Sendo desse modo implicando em algumas mudanças de estrutura no conjunto com adições de mais tabelas e exclusões/adições de colunas nas tabelas para fazer as novas alterações e ligações. Assim, para a conclusão do JusBD fazer os *replace* para modificar os tipos de atributos das colunas.

¹ CNJ: <http://www.cnj.jus.br/transparencia>

² QlikView: <https://www.qlik.com/>

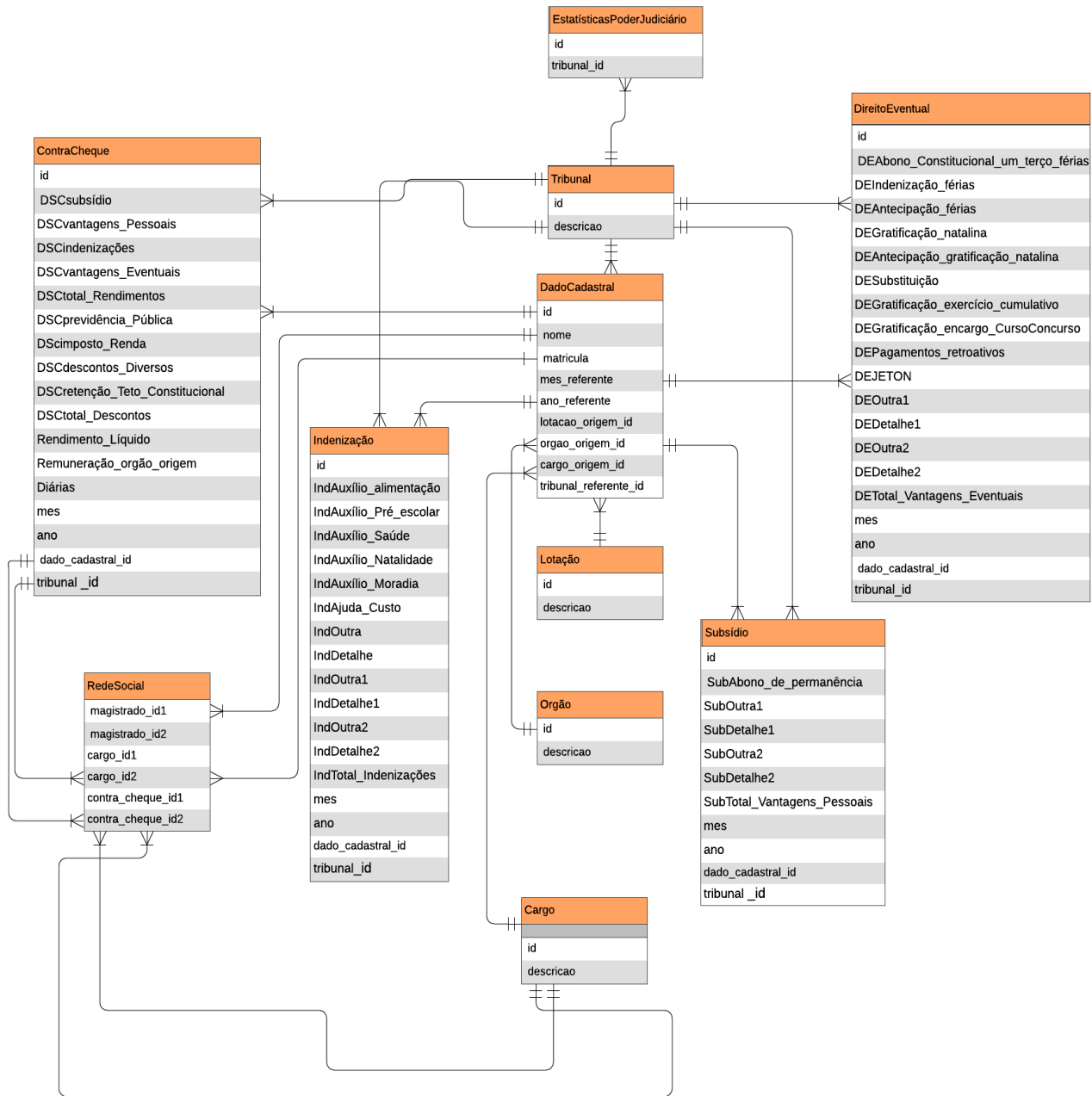


Figura 1 Modelagem relacional do JusBD.

Ademais, para melhorar o potencial do JusBD, inserimos a tabela RedeSocial, a ideia dela é conectar os magistrados que trabalharam no mesmo local e no mesmo período. Por exemplo, inserir uma linha na tabela para magistrados que trabalharam no mesmo local e possuem contracheque com ano no mesmo período. Assim, possibilitando uma conexão com os IDs entre dois magistrados formando uma tupla, sendo esses IDs chave estrangeira da tabela DadoCadastral.

RESULTADOS E DISCUSSÕES:

A pesquisa iniciou em abril de 2019 e está em andamento. Por isso, ainda não há resultados para serem apresentados.

CONCLUSÕES:

Durante o processo da construção do JusBD, houveram alguns desafios. Como empecilhos desse projeto pode-se dizer que foi a organização de tempo para aprendizagem e a falta de experiência em construção de um conjunto de dados, com o banco de dados trazendo algumas duplicidades e algumas perdas de informações específicas que serão resolvidas ao decorrer do tempo, por ser uma primeira versão. O aprendizado desse projeto foi desde desenvolver na prática um banco de dados (sem ter muito conhecimento sobre a área, assim aprendendo e aplicando) e todos seus processos e maneiras alternativas de manipular um conjunto de dados, até mesmo tipos de arquivos que tinha desconhecimento como csv e conjuntos como Gov Docs, T5 file corpus, entre outros.

REFERÊNCIAS BIBLIOGRÁFICAS:

[Anderson and Williams 2018] Anderson, S. and Williams, T. (2018). Cybersecurity and medical devices: Are the iso/iec 80001-2-2 technical controls up to the challenge? *Computer Standards & Interfaces*, 56:134–143.

[Becker et al. 2017] Becker, I., Hutchings, A., Abu-Salma, R., Anderson, R. J., Bohm, N., Murdoch, S. J., Sasse, M. A., and Stringhini, G. (2017). International comparison of bank fraud reimbursement: customer perceptions and contractual terms. *Journal of Cybersecurity*, 3(2):109–125.

[Dadkhah et al. 2018] Dadkhah, M., Lagzian, M., and Borchardt, G. (2018). Identity theft in the academic world leads to junk science. *Science and engineering ethics*, 24(1):287–290.

[Greengard 2012] Greengard, S. (2012). On the digital trail. *Communications of the ACM*, 55(11):19–21.

[Guarino 2013] Guarino, A. (2013). Digital forensics as a big data challenge. In *Procs. of ISSE securing electronic business processes*, pages 197–203. Springer.

Participação em Congressos, publicações e/ou pedidos de proteção intelectual:

Um artigo está sendo escrito para submissão no DSW (Dataset Showcase) do 34º Simpósio Brasileiro de Banco de Dados.