DRE-CVM: Explorando Dados Financeiros Enriquecidos com Proveniência e Princípios FAIR de Companhias Brasileiras de Capital Aberto

Gilberto Gil F. Gomes Passos[1], **Saulo A. Almeida**[1], **Valquire da S. de Jesus**[1], **Jorge Zavaleta**[1], **Sérgio Manuel Serra da Cruz**[1]

[1]Programa de Pós-Graduação em Informática (PPGI)
Universidade Federal do Rio de Janeiro (UFRJ)

Abstract. *Since 2005, the Organization for Economic Cooperation and Development (OECD) has encouraged the implementation of Financial Education (FE) initiatives for young people in various countries. The lack of FE can lead to financial difficulties for individuals, families, and businesses. With the aim of democratizing access to FE and promoting reflections on Economics and Finance for high school students throughout the country, the Brazilian Investment Olympiad (OBInvest) was created. This work presents the computational strategy DRE-CVM, which provides curated data series for OBInvest. Its architecture includes reproducible pipelines in container environments that generate high quality datasets, FAIRified and annotated with provenance metadata on financial information from the Securities and Exchange Commission.*

Resumo. *Desde 2005, a Organização Econômica de Cooperação e Desenvolvimento (OCDE) incentiva a implementação de iniciativas de Educação Financeira para jovens em diversos países. A falta de Educação Financeira (EF) pode levar a dificuldades financeiras para indivíduos, famílias e empresas. Com o objetivo de democratizar o acesso à EF e promover reflexões sobre Economia e Finanças para alunos do ensino médio em todo o país, foi criada a Olimpíada Brasileira de Investimento (OBInvest). Este trabalho apresenta a estratégia computacional DRE-CVM, que fornece séries de dados curados para a OBInvest. A estratégia inclui pipelines reprodutíveis em ambientes de contêineres que geram datasets de qualidade, FAIRificados e anotados com metadados de proveniência sobre informações financeiras da Comissão de Valores Mobiliários.*

1. Introdução

Desde 2005, a OCDE incentiva a implementação de iniciativas de Educação Financeira (EF) para jovens em diversos países. A falta de EF pode levar a dificuldades financeiras para indivíduos, famílias e empresas. Estudos da OCDE indicam que o brasileiro tem baixa EF e limitada cultura de investimentos. Iniciativas de EF e olimpíadas de conhecimentos visam mitigar esses problemas. De acordo com a BNCC do MEC [BNCC 2018], as escolas precisam ter a EF como tema em suas grades. Contudo, não precisa ser uma disciplina completa, o tema deve aparecer como assunto transversal em outras disciplinas, projetos ou competições. No Brasil, a OBInvest surgiu em agosto de 2020 no CEFET-RJ visando atender a BNCC, mitigar os problema da baixa EF e promover reflexões sobre questões econômicas e financeiras entre estudantes do nível médio.

A OBInvest estimula os estudantes pensarem em situações hipotéticas e tomarem decisões sobre investimentos. No entanto, ao longo das últimas edições da olimpíada, os autores verificaram que novas ferramentas educacionais poderiam ser agregadas ao processo ensino-aprendizagem envolvendo o (re)uso de *datasets* de qualidade sobre dados financeiros enriquecidos com metadados de proveniência [PROV-Overview 2013] e alinhados com os princípios FAIR [Wilkinson et al. 2016]. Estima-se que eles poderiam agregar novas perspectivas analíticas aos participantes visto que dados de melhor qualidade agregam mais confiabilidade às tomadas de decisões e menos riscos aos investidores.

Este trabalho apresenta uma estratégia computacional denominada DRE-CVM capaz de disponibilizar séries de dados financeiros curados, FAIRificados e anotados com metadados de proveniência, sobre demonstrações de empresas brasileiras de capital aberto presentes na Comissão de Valores Mobiliários (CVM). As demais seções do artigo estão organizadas da seguinte forma: a seção 2 discute os trabalhos relacionados. A seção 3, apresenta os materiais e métodos. A seção 4 apresenta a DRE-CVM. Na seção 5, se discutem os experimentos. A discussão é apresentada na seção 6. Por fim, a seção 7, apresenta as conclusões, limitações e sugestões para trabalhos futuros.

2. Trabalhos Relacionados

Atualmente, inexistem trabalhos na literatura nacional na área de ciência de dados que visam enriquecer com metadados de proveniência e alinhar os *datasets* da CVM com os princípios FAIR. [Chiella and Richartz 2019], investigam a geração e distribuição de riqueza criada por empresas listadas na CVM, com dados da Demonstração do Valor Adicionado (DVA). [Catapan and Colauto 2020] fazem análises comparativas da qualidade dos lucros de duas normas contábeis (COSIF e CPC) usando *datasets* da CVM e Banco Central do Brasil.

Existem projetos *open-source* disponíveis no Github que utilizam *datasets* da CVM, destaca-se [Quant 2020], que contém *notebooks* em Python para obtenção de variadas informações financeiras. O [Vido 2020] disponibiliza *notebooks* em Python que produzem dados das Demonstrações do Resultado do Exercício (DREs), *dataset* assemelhado ao analisado neste trabalho.

A proposta DRE-CVM se diferencia dos demais trabalhos tanto no métodos quanto na sua finalidade, é voltada para aprimorar a EF pois recomenda usos de dados abertos de companhias nacionais e é baseada em *pipelines* reprodutíveis em ambientes de contêineres produzindo *datasets* curados e alinhados aos princípios FAIR. Os princípios FAIR [Wilkinson et al. 2016], um acrônimo para *Findable, Accessible, Interoperable e Reusable*, estão presentes nas discussões da área de Ciência de Dados, são princípios norteadores aplicados na gestão de objetos digitais, em especial em dados científicos. Os princípios quando aplicados na gestão de dados científicos melhoram a qualidade dos dados e, consequentemente, sua interoperabilidade e capacidade de reuso.

3. Materiais e Métodos

Adotou-se a metodologia OSEMN [Mason and Wiggins 2010] como estratégia de Ciência de Dados e método de investigação nos processos analíticos. Nesta pesquisa, foram desenvolvidas *pipelines* e *workflows* para tratamento de dados das etapas de processos ETL, produzindo *datasets* curados e FAIRificados para a OBIvenst. A execução

dos experimentos computacionais se apoiou em ferramentas baseadas no Python3, biblioteca Pandas, plataforma KNIME, e em ambiente Jupyter, executando sobre Anaconda3 em contêineres Docker. Parte dos experimentos foram executados no ambiente de nuvem Google Colaboratory. Os dados brutos são abertos e oriundos da CVM (`https://dados.cvm.gov.br/dataset/`), dispostos em formato CSV; eles são organizados em três *datasets* : ***(i)*** Dados Cadastrais de Companhias Abertas (CAD-CIA), total de 2.550 registros), ***(ii)*** 1.654.787 registros do formulário de Informações Trimestrais (ITR) e ***(iii)*** 665.666 registros do formulário de Demonstrações Financeiras Padronizadas (DFP).

4. DRE-CVM

Como contribuição, expandiu-se a arquitetura DRE-CVM [Almeida et al. 2023], que se baseia em uma estratégia computacional que produz séries de dados curados. Ela está conceitualmente representada na Figura 1, é composta por *pipelines* e *workflows* reprodutíveis em ambientes de contêineres que produzem *datasets* curados, FAIRificados e anotados com metadados de proveniência retrospectiva [Cruz et al. 2009] sobre as demonstrações financeiras de empresas brasileiras de capital aberto presentes na CVM.

4.1. Pipelines ETL

A preparação e limpeza de dados brutos foi mediada por *pipelines* baseados na biblioteca Pandas. Foram processados os *datasets* CAD-CIA, ITR e DFP e realizada a fusão para oferecer novos *datasets* curados que atendessem as necessidades da disseminação dos esforços educacionais de difusão de EF. As principais atividades de preparação dos dados incluíram a transformação, filtragem de empresas com cadastro ativo e agregação e fatiamento dos *datasets* ITR e DFP para obtenção de informações consolidadas do DRE, através da subtração de valores e concatenação de *dataframes*.

A fase da exploração de dados foi conduzida no ambiente KNIME através de *workflows* exploratórios dos *datasets*. Identificaram-se as operações necessárias para a FAIRificação, anotação e criação do *dataset* curados utilizáveis nas próximas edições da OBInvest. Os códigos-fonte do *workflow* KNIME e *pipelines* estão disponível no repositório do projeto.

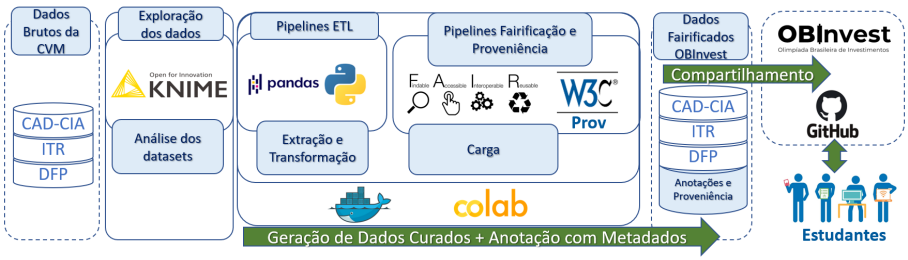

Figura 1. Representação Conceitual do DRE-CVM.

5. Experimentos Computacionais

Essa seção apresenta os experimentos de FAIRificação de dados e construção dos grafos de proveniência e de verificação de reprodutibilidade na DRE-CVM. O repositório

do projeto foi disponibilizado no Github, ele contém os artefatos produzidos no DOI registrado no Zenodo, em: `https://doi.org/10.5281/zenodo.7110653`. Essa estratégia permite que pesquisadores tenham livre acesso ao conteúdo e possam acessá-los e reproduzi-los através de contêineres Docker.

Os *datasets* curados do OBInvest contém os DREs trimestrais das empresas de capital aberto da CVM, sendo um dos resultados dos experimentos. Procurou-se seguir as práticas relacionadas à FAIRificação de dados. Além disso, há uma descrição mais detalhada dos experimentos, da proveniência com vistas a ampliar a interoperabilidade, reusabilidade de dados e reprodutibilidade dos experimentos por terceiros.

5.1. FAIRificação de Dados

A chamada FAIRificação de dados (do inglês *FAIRification*) deve ser aplicada sobre todo o ciclo de vida de dados e metadados do experimento, para que se estes se tornem aderentes aos Princípios FAIR [Wilkinson et al. 2016]. O processo de FAIRificação adota um modelo semântico condizente com os dados experimentais, sobre a forma como a gestão de dados pesquisas devem ser conduzidas. Neste trabalho, a FAIRificação ocorreu através dos *workflows*, por exemplo, na geração de *ids* globais e persistentes para os *datasets* curados para facilitar a sua localização por humanos ou máquinas. Também trabalhou-se na definição de metadados e produção de *datasets* estruturados usando padrões de dados abertos na combinação/agregação com outros *datasets*; na geração de metadados de proveniência e acesso e licenças de uso aberto dos *datasets* para possam ser replicados ou reutilizados em diferentes pesquisas ou iniciativas de EF. Devido ao escopo do planejamento e da execução dos experimentos e da ausência de repositórios públicos do tipo *FAIR Data Point* (FDP) nesta fase da pesquisa, optou-se por alinhar os *datasets* aos princípios de encontrabilidade e reuso.

5.2. Proveniência dos *Datasets*

Os metadados de proveniência retrospectiva representam, em detalhes, a execução de um experimento científico: desde a origem e a localização dos dados utilizados; todos os agentes envolvidos; e um passo a passo da execução do experimento, onde os metadados que permitem a interoperabilidade. A proveniência dos experimentos foi coletada durante a execução dos *pipelines* e utilizou-se a biblioteca PROV, baseada no padrão PROV da W3C.

Resumidamente, a proveniência dos experimentos pode ser visualizada em três partes principais. A primeira parte, indica a origem dos dados brutos, disponibilizados pelo agente CVM, a segunda parte apresenta a hierarquia de agentes envolvidas na definição e execução dos experimentos. Finalmente, na terceira parte, foi criado um agente de software referente ao *notebook* python do experimento, que detalha toda a execução do experimento, incluindo os contêineres e *timestamps* de cada etapa durante a sua execução. Os grafos de proveniência dos experimentos estão acessíveis no repositório deste projeto.

5.3. Reprodutibilidade dos Experimentos

Um ponto importante abordado neste pesquisa foi a assegurar a repetibilidade e reprodutibilidade dos experimentos computacionais. Embora ambientes em nuvem como o Google

Colaboratory ou o MyBinder sejam opções atraentes, eles podem sofrer mudanças independentemente da vontade dos pesquisadores; por exemplo, trazendo preocupação sobre à evolução das linguagens e de suas bibliotecas e como isso poderia afetar a futura repetibilidade dos mesmos. Para mitigar essas questões, a estratégia DRE-CVM apoiou-se em contêineres Docker em ambiente Python gerenciado pelo Conda, que permitiu controlar as versões dos *pipelines*, dados e bibliotecas utilizadas.

Durante a criação da imagem Docker, foram fixadas as versões do Python e suas bibliotecas. Também foi desenvolvido uma abordagem para realizar a verificação dessas versões antes de executar os códigos dos experimentos computacionais. Além disso, clonou-se uma versão específica do repositório do OBInvest (usando *tags* de controle de versão) para o diretório que serve como *path* para a instância do Jupyter executada durante a execução da imagem. Isso permitiu reproduzir todos os experimentos, incluindo a criação dos *datasets* da OBInvest e registro de metadados e grafos de proveniência em tempo de execução.

A imagem Docker foi compartilhada no Docker Hub, repositório remoto de imagens Docker, que permite baixar e executar um contêiner de uma imagem de forma simplificada mesmo por usuários pouco experientes. Devido a limitações de espaço neste artigo, os detalhes da execução dos experimentos estão detalhados no repositório do projeto.

6. Discussão

Até o momento, não se conseguiu localizar trabalhos relacionados que tratem os dados dos casos excepcionais que ocorrem no balanço financeiro trimestral da CVM. Embora fossem detectados esses casos e excluído dos *datasets* curados, acredita-se que o conjunto de dados curados será útil para as atividades de EF da OBInvest.

Apesar da falta de referências na literatura que correlacionam os temas EF, FAIRificação de dados e reprodutibilidade de experimentos usando contêineres, considera-se que elaboração da estratégia computacional foi bem-sucedida ao se considerar a elevada reprodutibilidade dos experimentos que podem ser conduzidos por equipes distintas usando um mesmo ambiente computacional oferecido pelo contêineres. Os procedimentos de geração de *datasets* curados podem ser facilmente executados pelas equipes organizadoras das futuras edições da OBInvest a partir de qualquer computador com conexão à Internet e após a primeira execução, pode até ser usado em modo ”off-line”em execuções subsequentes.

Os metadados de proveniência retrospectiva gerados durante o desenvolvimento do pesquisa fornecem descritores detalhadas sobre a origem dos dados brutos, os principais agentes envolvidos nas transformações de dados e o passo a passo dos experimentos computacionais. Essas informações são criadas em tempo de execução nos formatos PNG, RDF Turtle e XML, para a ampliar a reusabilidade dos dados. As análises de dados de margem do setor, embora simples, ilustram claramente o potencial de informações que podem ser obtidas a partir do conjunto de dados curados criados nos experimentos.

7. Conclusão

Este artigo apresenta uma estratégia para enriquecer com metadados de proveniência os *dataset*s da CVM e produzir *datasets* curados sobre demonstrações financeiras trimestrais de empresas de capital aberto para uso na OBInvest. Como exemplo do potencial

dessa estratégia, foram disponibilizados todos os artefatos computacionais e uma análise exploratória de margem trimestral agrupada por setor industrial, tais demonstrativos estão disponíveis no repositório do projeto.

Durante a pesquisa, foram incorporados os princípios de FAIRificação de dados e empacotamos o *design* dos *pipelines*, *workflows* e o ambiente utilizado em imagens Docker para auxiliar a reprodutibilidade dos experimentos por partes interessadas em EF ou atividades afins.

Melhorias podem ser feitas futuramente, a saber: ***(i)*** realizar análises financeiras mais profundas nos *datasets*, como prever resultados com base no histórico de demonstrações financeiras; ***(ii)*** adaptar a etapa de verificação da versão dos componentes para comparar a lista completa extraída do ambiente Conda com o ambiente identificado no momento da execução do experimento; ***(iii)*** aprofundar a etapa de FAIRicação de dados e investigar conexões dos *datasets* com repositórios semânticos do tipo FDP.

Referências

Almeida, S., Passos, G., Jesus, V., Serra, S., and Zavaleta, J. (2023). Providing data on financial results of public companies enriched with provenance for obinvest. In *ACM Web Conference WWW ´2023*, page 1563, New York, NY, USA.

BNCC (2018). Base nacional comum curricular. Disponível em: `http://basenacionalcomum.mec.gov.br/images/BNCC_EI_EF_110518_versaofinal_site.pdf`, Acessado em: 20/09/2022.

Catapan, A. and Colauto, R. D. (2020). Governança corporativa: uma análise de sua relação com o desempenho econômico-financeiro de empresas cotadas no brasil nos anos de 2010–2012. *Contaduría y administración*, 59(3):137–164.

Chiella, F. and Richartz, F. (2019). Mfc283-análise da geração e distribuição do valor adicionado das empresas registradas na cvm nos anos de 2009 a 2017.

Cruz, S. M. S. d., Campos, M. L. M., and Mattoso, M. (2009). Towards a taxonomy of provenance in scientific workflow management systems. In *2009 Congress on Services - I*, pages 259–266.

Mason, H. and Wiggins, C. (2010). A taxonomy of data science. Disponível em: `https://https://web.archive.org/web/20211219192027/http://www.dataists.com/2010/09/a-taxonomy-of-data-science/`, Acessado em: 26/09/2022.

PROV-Overview (2013). An overview of the prov family of documents. Disponível em: `https://www.w3.org/TR/prov-overview/`, Acessado em: 12/10/2022.

Quant, C. (2020). Python para investimentos. Disponível em: `https://github.com/codigoquant/python_para_investimentos`, Acessado em: 12/10/2022.

Vido, L. (2020). Dre cvm. Disponível em: `https://gist.github.com/Vido/cbc33862dd27a22790df633f1d113ae6`, Acessado em: 12/10/2022.

Wilkinson, M. D. et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.