Processo de ETL para inserção de dados estatísticos do ensino superior brasileiro com foco em mulheres nas áreas STEM no grafo de conhecimento da plataforma ELLAS

Bruna Oenning Amador, Rita Cristina Galarraga Berardi

Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, PR – Brasil

brunaamador@alunos.utfpr.edu.br, ritaberardi@utfpr.edu.br

Abstract. *An efficient discussion about the under-representation of women in STEM areas in Latin America must be based on data. Therefore, this article proposes an ETL process adapted to Linked Data for the integration of statistical data from INEP focusing on women in ELLAS platform's knowledge graph. A pipeline was implemented that answers questions about women in Brazilian higher education in STEM areas. This work contributes to the field of knowledge graphs by exposing a practical application and its developments.*

Resumo. *Uma discussão eficiente sobre a sub-representação de mulheres em áreas STEM na América Latina deve ser baseada em dados. Posto isso, este artigo tem como objetivo propor um processo de ETL adaptado para Dados Conectados para a integração de dados estatísticos do INEP com foco em mulheres no grafo de conhecimento da plataforma ELLAS. Foi implementado um pipeline que responde questões sobre sobre mulheres na educação superior brasileira em áreas STEM. Este trabalho contribui para a área de grafos de conhecimento por expor uma aplicação prática e seus desdobramentos.*

1. Introdução

Historicamente, mulheres latino-americanas têm sido sub-representadas nas ocupações em Ciência, Tecnologia, Engenharia e Matemática (STEM) e são menos propensas a iniciarem e persistirem em cursos de graduação nestas áreas [Granovskiy 2018]. A fim de identificar os fatores que influenciam esta problemática, é necessário coletar e analisar diversos dados, entretanto, estes dados geralmente não são disponibilizados em qualidade e formato próprios para reuso [Hildebrand et al. 2024]. Deste modo, a repetição de esforços para coletar e tratar esses dados se torna comum, o que dificulta a colaboração e novas contribuições, resultando em pesquisas isoladas. Ainda que estes dados sejam disponibilizados de forma aberta e em formatos que auxiliam no reuso, há o desafio em como integrar diferentes formatos e fontes de dados, de modo que se tenha uma visão sistêmica destes dados considerando o contexto latino americano. Neste sentido, a rede de pesquisa ELLAS foi criada para contribuir na geração de grafos de conhecimento para uso de dados em formato aberto e comparáveis entre países [Maciel et al. 2023], a fim de evidenciar e reduzir a diferença de gênero STEM na América Latina, com universidades parceiras do Brasil, Peru e Bolívia. A coleta de dados tem sido organizada em dois tipos de dados: Dados primários: dados não estruturados, provenientes de artigos científicos e até mídias sociais e coletados por meio de pesquisas como Revisão Sistemática da Literatura; e dados secundários: dados semi estruturados provenientes de sites de organizações nacionais e internacionais, em geral de portais de dados abertos [Hildebrand et al. 2024]. O presente trabalho atua na coleta de dados

secundários provenientes do Censo da Educação Superior[1] realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), órgão federal brasileiro responsável por indicadores educacionais. Este Censo disponibiliza anualmente dados que podem ser usados e compartilhados livremente sobre instituições de educação superior (IES), estudantes e docentes. Entretanto, há o desafio da interoperabilidade destes dados, como "conversar" informações em formatos distintos e poder compará-las de modo que representem um mesmo significado.

O presente trabalho é uma extensão da pesquisa de Souza (2023) que iniciou a criação de um modelo ontológico para integração dos dados do INEP do ano de 2021 com o recorte de gênero para a computação. No entanto, foram identificados alguns pontos de melhorias. Tendo em vista que todo o processo desenvolvido por Souza (2023) foi realizado de forma manual, surge a necessidade de que estes dados possam ser atualizados anualmente na plataforma ELLAS de forma automatizada, contemplando então uma uma marcação temporal a ser definida. Também, a utilização somente de dados de cursos da área da computação não são representativos, sendo preciso englobar as áreas de STEM, e propor um maior enriquecimento semântico da modelagem, adequando às boas práticas, como o reuso de ontologias e desvincular a ontologia do dicionário de dados disponibilizado pelo INEP. Tendo isso em vista, o presente artigo tem como objetivo propor um processo automatizado via ETL (Extração, Transformação e Carga) para triplificação de dados estatísticos do INEP com foco em mulheres nos cursos de áreas STEM e integração no grafo de conhecimento da plataforma ELLAS. Quanto à metodologia, a pesquisa é aplicada descritiva [Gil 2022], porque descreveu características de mulheres em STEM no contexto da Educação Superior brasileira e analisou a existência de associações entre variáveis como a presença de mulheres em diferentes cursos nas áreas STEM.

2. Trabalhos Relacionados

Em Costa et al. (2022), discute-se a integração de dados de sensores utilizando grafos do conhecimento, e como os diferentes níveis de dados IoT são incorporados nesses grafos, representados pela hierarquia de Dados, Informação, Conhecimento e Sabedoria. A utilização de grafos de conhecimento traz benefícios em diversos contextos, porque permite desenvolver serviços mais especializados e proporciona uma visão ampliada para a tomada de decisões [Costa et al. 2022]. Esses fatores são positivos e relevantes para a problemática abordada no presente artigo. Santos (2016), Cordeiro et al. (2011), Magalhães e Cardoso (2016), e Rodrigues e Maciel (2022) desenvolveram um processo ETL, utilizando Pentaho *Data Integration* (PDI), para trabalhar com dados abertos de diferentes fontes governamentais e educacionais do Brasil. O enfoque desses autores é a integração destes dados, para transformá-los em formatos como *Data Warehouse* (DW) e RDF. As principais diferenças entre estes trabalhos está nos contextos específicos de aplicação: Santos (2016) e Cordeiro et al. (2011) trabalharam com dados do governo brasileiro, mas não provenientes do INEP; Magalhães e Cardoso (2016) analisam dados do ensino superior brasileiro (INEP) no contexto de DW; e Rodrigues e Maciel (2022) se concentram em dados educacionais da Universidade Federal de Mato Grosso.

[1]www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior

3. ETL em Dados Abertos Conectados

Berners-Lee (2006) propôs um conjunto de práticas recomendadas para a publicação de dados estruturados na Web, fundamentado no conceito de Dados Abertos Conectados. Dados Abertos referem-se a dados que qualquer pessoa pode acessar, usar, modificar e compartilhar livremente para qualquer finalidade, estando disponível na Web em um formato compreensível por máquina [Open Knowledge 2024]. Berners-Lee (2006) discute que esses Dados Abertos se tornam progressivamente mais poderosos quando conectados, posto que há a criação de *links* entre os dados, que podem ser explorados por uma pessoa ou máquina. Deste modo, com o uso de Dados Conectados, é possível encontrar outros dados relacionados dentro dessa teia de dados, conhecida como grafos de conhecimento. Para isso utiliza-se *Uniform Resource Identifier* (URI) para nomear as coisas. Isso permite a busca por relações entre dados provenientes de fontes distintas, fornecendo contexto e possibilitando a descoberta de novos conhecimentos. A integração destes dados se dá pela utilização de ontologias, uma forma de representação do conhecimento, a fim de definir um vocabulário comum sobre os dados [Le-Phuoc et al. 2016]. Para representar esses recursos na Web utiliza-se o *Resource Description Framework* (RDF). No entanto, o processo de publicação de Dados Conectados e a suas operações de extração, preparação, modificação, triplificação, e integração, tem carência por um processo que facilite e oriente os usuários quanto a esses passos [Santos 2016].

Neste sentido, para que seja possível a inserção de dados de fontes externas para um grafo de conhecimento, os dados precisam passar por um processo de ETL (Extrair, Transformar e Carregar), permitindo a centralização das operações realizadas e aumentar a facilidade de acesso dos dados. Um processo de ETL busca extrair dados dos sistemas de origem, impondo padrões de qualidade e consistência de dados, para que fontes independentes possam ser usadas em conjunto e, por fim, fornecer dados em um formato padronizado e pronto para utilização [Kimball e Caserta 2004]. Contudo, um processo de ETL foi originalmente projetado com foco em projetos de *Data Warehouse* (DW), sendo preciso encontrar soluções para adequar este processo de ETL para Dados Conectados. Isto porque existem semelhanças entre ambos os processos de publicação, como a extração de dados úteis de diferentes fontes existentes e os dados são convertidos de um formato para outro. Logo, os benefícios imediatos da abordagem ETL para Dados Conectados, segundo Santos (2016), são: sistematização do processo de publicação de Dados Conectados; monitoramento e gerenciamento das atividades de extração, transformação e carga; oportunidade de reutilização do *workflow* para carregar novos dados e atualizar dados previamente publicados.

4. Modelagem da Ontologia

A ontologia foi estendida a partir da primeira versão da ontologia ELLAS, criada para o contexto de dados primários preexistentes, que contém dados de Iniciativas e Políticas voltadas para a inserção e manutenção de mulheres em carreiras STEM, e Fatores que influenciam a presença ou ausência de mulheres nessas carreiras. A classe localização é o principal eixo de conexão entre os contextos de dados já existentes no grafo de conhecimento com os dados do INEP, onde ambos os contextos se conectam, gerando a versão ELLAS 2.0 da ontologia. Para os dados secundários sobre educação superior, foi necessário um estudo detalhado dos dados disponibilizados pelo INEP, de modo a

extrair e representar ao máximo a semântica dos dados. Duas principais classes foram mapeadas: Curso e Instituição de Ensino Superior (IES). A partir delas, foram identificadas restrições que fizessem sentido no contexto do INEP. Desta forma, foram estabelecidas relações de subclasse utilizando classificações fornecidas pelo INEP, como o tipo de organização acadêmica da IES e o tipo do grau acadêmico do curso. Essas classificações são disjuntas, ou seja, só podem pertencer a uma subclasse ao mesmo tempo. É importante destacar que não foram identificadas ontologias para reuso que pudessem auxiliar na modelagem, visto que o INEP possui um escopo fechado. Tendo em vista que este trabalho está inserido em um contexto latino-americano, é preciso considerar a escalabilidade da modelagem e tratar diferenças culturais destes países, de modo que a ontologia esteja apta a representar não só os dados da educação superior brasileira, mas também possa receber dados deste contexto sobre outros países. Neste sentido, foi identificado que o INEP trabalha com uma classificação, chamada de Cine Brasil[2], que padroniza as nomenclaturas dos cursos e as classifica em quatro níveis. Esta classificação tem como base a classificação CINE (Classificação Internacional Normalizada da Educação) da UNESCO, que busca facilitar comparações de sistemas educativos entre países [UNESCO 2024] e pode ser utilizada para padronização dos dados de diferentes países. Também, para assegurar que o significado pretendido de cada termo da ontologia seja claro e preciso, foi utilizada uma marcação de linguagem, inicialmente para o português e inglês, língua comum adotada para comunicação dentro do projeto, facilitando a colaboração e integração latino-americana.

Para a extensão da ontologia, também foi preciso pensar em como trazer uma marcação temporal, porque existe a necessidade em saber sobre qual ano os dados se referem. Os grafos do conhecimento não foram originalmente projetados para lidar com dados temporais e ainda não há uma forma padronizada e sofisticada de resolver este problema. Foram aprofundadas abordagens que buscam alternativas para tratar o problema da temporalidade e, levando em conta a complexidade e escalabilidade da aplicação, foi escolhida a marcação temporal utilizando *named graphs*. A ideia desta abordagem é que as representações temporais se dão por meio de nomear os grafos para poder falar algo sobre eles. Assim, é possível atribuir um nome a um grafo RDF, utilizando um identificador na forma de um URI [Carroll et al. 2005], para que este identificador possa ser tratado como um nó no grafo RDF, possibilitando fazer declarações sobre todo o *named graph*. Essa abordagem resultou em uma solução melhor estruturada, tanto na apresentação quanto na semântica dos dados. Com relação às buscas no grafo de conhecimento, é possível dizer sobre qual *named graph* a busca deve ser realizada, além de poder trabalhar com mais de um *named graph* em conjunto.

5. Processo de ETL

Para ser um processo auto sustentável ao longo do tempo, relativamente independente da fonte de subsídio que mantém o engajamento da equipe do projeto, se torna necessário automatizar este processo de ETL dos dados secundários de modo que a cada atualização da fonte de dados os dados do projeto também sejam atualizados [Hildebrand 2024]. A Figura 1 apresenta a implementação do processo automatizado de

[2] https://www.gov.br /inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/cine-brasil.

ETL no *Pentaho Data Integration* (PDI)[3] para o contexto de dados do INEP. A primeira etapa deste processo é a extração de dados úteis de diferentes fontes existentes. A segunda etapa é a transformação, na qual as inconsistências dos dados são eliminadas e os dados são convertidos de um formato para outro [Cordeiro et al. 2011]. Por fim, os dados são carregados para o grafo de conhecimento da plataforma ELLAS. Foram utilizadas diversas ferramentas para o desenvolvimento, sendo elas: OntoRefine, Protègè, GraphDB, Python e PDI. O processo de ETL foi desenvolvido utilizando PDI, por apresentar um bom desempenho na manipulação de grande volumes de dados e por ter uma interface que auxilia na visualização do *pipeline*. Um ponto positivo do INEP é a padronização ao longo dos anos, que facilita a extração e transformação dos dados. Primeiramente, na etapa de extração (1), foi chamado pelo *pipeline* um código python. Neste código, é verificado se há uma nova atualização do INEP e se o padrão de colunas dos dados foi alterado, crucial para o mapeamento dos dados. A extração ocorre por meio de um *web scraping* que busca todos os *links* da página, que possibilita realizar o *download* dos arquivos ZIP disponíveis, carregados na pasta INEP no diretório. Esta etapa também delimita o ano mínimo que os dados serão extraídos e quais dados já foram carregados no grafo de conhecimento.

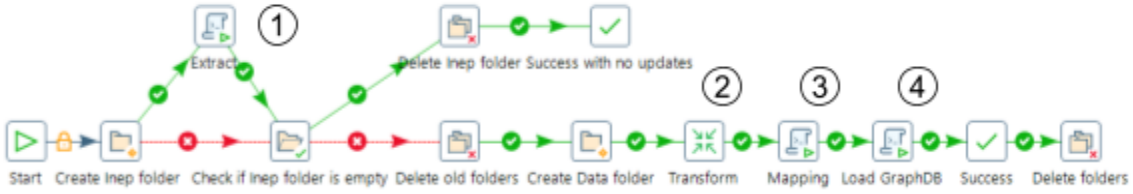

Figura 1. Processo automatizado de ETL (Extrair, Transformar e Carregar)

A etapa de transformação se dá em duas principais etapas, a limpeza dos dados (2), utilizando PDI, e o mapeamento (3), que transforma os dados do formato CSV para triplas RDF via código python. Para a limpeza dos dados, foi necessário realizar um *loop* no PDI para ser possível passar mais de um arquivo para o *pipeline*, considerando que mais de um ano pode estar sendo carregado no grafo de conhecimento. Tanto para os cursos quanto para as IES, foram aplicadas as seguintes etapas: Seleção das colunas que melhor se enquadram ao contexto do ELLAS e quais foram necessárias para realizar o mapeamento dos dados; Remoção de caracteres especiais; Substituição dos valores do dicionário de dados; Remoção de espaços indesejados; Organização dos códigos da IES por ordem crescente; Junção dos dados de cursos e IES em um único arquivo CSV. Destaca-se as etapas de filtrar as linhas e a manipulação do código identificador que são exclusivas dos cursos. Primeiramente, ao filtrar as linhas dos cursos é definido o que será considerado como cursos de áreas STEM. Este filtro foi realizado com base na classificação Cine Brasil, buscando maior padronização e descendo nos níveis caso o resultado não seja satisfatório. Para julgar estes resultados foram baseados em Koonce et al. (2011), que define STEM na perspectiva educacional. Em seguida, a concatenação do código do curso com os códigos de cidade e estado foi necessário para diferenciar onde que cada curso está localizado, posto que a IES contém somente a localização da sede matriz. Para o mapeamento, criado previamente de forma manual, foi aplicado via

[3] A versão utilizada foi a *Pentaho Data Integration Community Edition* 9.3. Mais informações em: https://pentaho.com/products/pentaho-data-integration/

linha de comando do Onto Refine para cada arquivo CSV resultante, gerando um arquivo RDF para cada ano. Foram utilizadas as colunas do CSV como referência para a triplificação dos dados, sendo de extrema importância que existam. Ao final, é alterado o arquivo CSV que controla os dados carregados no grafo de conhecimento.

Por fim, na etapa de carga (4), foi chamado um arquivo python que abre o diretório com os arquivos RDF e, para cada arquivo existente, busca por uma linha que contenha **SecondaryData**, isso porque é a identificação dos *named graphs* que será passada como parâmetro "*context*" para o GraphDB[4], que será responsável pela carga. Os dados são carregados no grafo de conhecimento com o uso de linha de comando do GraphDB. Vale ressaltar que é preciso criar o repositório manualmente no GraphDB e adicionar o *default graph*[5] diretamente no GraphDB, pois não devem ser atualizados pelo *pipeline*. O *pipeline* foi executado com o carregamento de dados de 2009 até 2022 e foram selecionadas 32 de 200 colunas (Cursos) e 17 das 84 colunas (IES). Percebeu-se com isso que há um aumento no tamanho dos arquivos do INEP ao longo dos anos, de 30,7 MB em 2009 para 318 MB em 2022. Como consequência, há aumento no tempo de execução do *pipeline*, que passa de 41 segundos em 2009 para cerca de 2 minutos em 2022. O tempo total de execução foi de 13 minutos e 5 segundos. A partir desta execução de sucesso foi possível responder Questões de Competência, tais como: "Qual a porcentagem de mulheres matriculadas nos cursos de Sistemas de Informação por estados brasileiros de 2019 a 2022?" e "Quantas mulheres concluíram bacharelados em cursos de engenharia em universidades públicas no Sul do Brasil em 2022?". O detalhamento dos testes, o *pipeline* e também as Questões de Competência com suas respectivas *queries* estão disponíveis no GitHub[6].

6. Conclusão

O presente artigo apresenta a implementação de um processo automatizado via ETL para triplificação de dados estatísticos do INEP com foco em mulheres nos cursos de STEM. Este processo contribui com o reuso destes dados, mantendo a interoperabilidade e estabelecendo um padrão de representação de dados de educação superior no contexto da América Latina, para permitir novas atualizações destes dados de modo automatizado para o grafo de conhecimento da plataforma ELLAS. Para isso, foi preciso adequar o processo de ETL para Dados Conectados e tratar a temporalidade destes dados em grafos do conhecimento, sendo possível adequar para outras bases de dados da educação superior. Após esta integração, foi possível responder Questões de Competência, que evidenciam a aplicabilidade e o poder do grafo de conhecimento para pesquisas e análises que buscam auxiliar na identificação de possíveis fatores para a sub-representação de mulheres em STEM. Como trabalhos futuros, busca-se trazer uma marcação de linguagem da ontologia para o espanhol, expandir a fonte de dados para incluir outros países latino americanos e implementar o *pipeline* no servidor do ELLAS.

[4] https://graphdb.ontotext.com/
[5] Grafo em que são importados os axiomas definidos pela ontologia.
[6] https://github.com/brunaoenning/TCC

Referências

Carroll, J. J., Bizer, C., Hayes, P., Stickler, P. (2005) Named graphs. Journal of Web Semantics, Elsevier, v. 3, n. 4, p. 247–267. ISSN 1570-8268.

Cordeiro, K. F., et al. (2011). An approach for managing and semantically enriching the publication of linked open government data. 3rd Workshop in Applied Computing for Electronic Government (WCGE). 82-95.

Costa, F., Avila, C., Rolim, T., Andrade, R., & Vidal, V. (2022). DIKW4IoT: Uma abordagem baseada na hierarquia DIKW para a construção de grafos de conhecimento para integração de dados de IoT. In Anais do XXXVII Simpósio Brasileiro de Bancos de Dados, (pp. 190-202). Porto Alegre: SBC.

Gil, A. C. (2022) Como elaborar projetos de pesquisa. [S.l.]: Editora Atlas. ISBN 9786559771646.

Granovskiy, B. (2018) Science, technology, engineering, and mathematics (stem) education: An overview. Congressional Research Service, ERIC.

Hildebrand, N., Amador, B., Maciel, C. e Berardi, R. (2024) A escassez de dados abertos estruturados em países latino-americanos com enfoque de gênero na educação superior. In Anais do XVIII Women in Information Technology, julho 21, 2024, Brasília/DF, Brasil. SBC, Porto Alegre, Brasil, 160-171.

Kimball, R., Caserta, J. (2011). The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Alemanha: Wiley.

Koonce, D., Zhou, J., Anderson, C., Hening, D. (2011) What is stem? ASEE Annual Conference Exposition, ASEE Conferences, Vancouver, n. 10.18260/1-2–18582.

Maciel, C., Guzman, I., Berardi, R., Caballero, BB, Rodriguez-Rodriguez, N., Frigo, L., Salgado, L., Jimenez, E., Bim, SA e Tapia, PC (2023) Plataforma de dados abertos para promover políticas de igualdade de gênero em STEM, em Anais do Western Decision Sciences Institute (WDSI). Abril de 2023. Portland Oregon, EUA.

Magalhães, H., Cardoso, L. (2016). Análise de dados abertos sobre o ensino superior brasileiro. Trabalho de conclusão de curso, Universidade de Brasília, Brasília - DF.

Open Knowledge (2024). Open Definition - Defining Open in Open Data, Open Content and Open Knowledge. Disponível em: https://opendefinition.org.

Phuoc, D., Quoc, H., Ngo, H., Nhat, T., Hauswirth, M. (2016). The Graph of Things: A step towards the Live Knowledge Graph of connected things. Web Semantics: Science, Services and Agents on the World Wide Web. 37. 25-35.

Rodrigues, F., & Maciel, C. (2022). Um método para captura e compartilhamento de dados abertos educacionais via um processo ETL. In Anais do X Workshop de Computação Aplicada em Governo Eletrônico, (pp. 133-144). Porto Alegre: SBC.

Santos, S. (2016) Um processo para conversão e publicação de dados para modelo rdf seguindo os princípios de Linked Data. TCC (Graduação em Sistemas de Informação) - Universidade Federal do Ceará, Campus Quixadá, Quixadá.

Souza, V. L. (2023) Elaboração de uma ontologia que endereça a presença de mulheres em cursos de computação no brasil.TCC (Graduação em Sistemas de Informação) - Universidade Tecnológica Federal do Paraná, Campus Curitiba.

UNESCO. (2024) International Standard Classification of Education. Disponível em: https://uis.unesco.org/en/topic/international-standard-classification-education-isced.