ELLAS: Uma plataforma de dados abertos com foco em lideranças femininas em STEM no contexto da América Latina

Rita Cristina Galarraga Berardi[1], Pedro Henrique Stolarski Auceli[1], Cristiano Maciel[2,4], Guillermo Davila[3], Indira R. Guzman[4], Luana Mendes[2,5]

[1]Departamento de Informática
Universidade Tecnológica Federal do Paraná (UTFPR) – Curitiba – PR – Brazil

[2]Instituto de Computação
Universidade Federal de Mato Grosso (UFMT) – Cuiabá – MT – Brazil

[3]Instituto de Investigación Científica – Faculdad de Ingeniería – Carrera de Ingeniería de Sistemas – Universidad de Lima (ULIMA) – Lima – Peru

[4]Computer Information Systems – College of Business Administration
California State Polytechnic University (CalPoly) – Pomona – California – USA

[5]Fundação de Apoio e Desenvolvimento da Universidade Federal de Mato Grosso (UNISELVA) – Cuiabá – MT – Brazil

ritaberardi@utfpr.edu.br, pedroauceli@alunos.utfpr.edu.br, cristiano.maciel@ufmt.br, gdavila@ulima.edu.pe, irguzman@cpp.edu, ellas.latinamerica@gmail.com

Abstract. *Latin American and Caribbean countries suffer from the underrepresentation of women in the STEM areas and to solve this problem initiatives, policies and actions are often created by different bodies, from either public or private sectors. However, both data and information on initiatives, policies and actions, are neither open, nor concentrated and sometimes not even structured. To reduce this problem, and to facilitate analysis research that uses data on women in STEM, the ELLAS project was created. The objective of the project is to create a platform with technologies based on the Semantic Web to structure and concentrate data from Brazil, Peru and Bolivia, initially. This paper will present the strategies utilized to develop this platform.*

Resumo. *Os países da América latina e Caribe sofrem com a sub-representação de mulheres nas áreas de STEM e para resolver esse problema iniciativas, políticas e ações são criadas com frequência por diversos órgãos, sejam eles públicos ou privados. Porém, dados e informações de iniciativas, políticas e ações, não estão abertos, concentrados em algum local e nem sempre estruturados. Para reduzir esse problema e facilitar pesquisas que utilizam dados sobre mulheres em STEM, a rede de pesquisa internacional ELLAS foi criada e executa um projeto. O intuito do projeto é a criação de uma plataforma com tecnologias baseada em Web Semântica para estruturar e concentrar os dados do Brasil, Peru e Bolívia, inicialmente. Assim, esse artigo tem por objetivo apresentar as estratégias adotadas para o desenvolvimento desta plataforma.*

1. Introdução

Refletindo sobre a situação da ciência, tecnologia e inovação na América Latina e no Caribe, fica claro que a economia desta região não está bem preparada para enfrentar os desafios da sociedade do conhecimento (Guzman et al., 2020). Parte do problema é que não há estudantes suficientes se engajando em *Ciências, Tecnologia, Engenharia e Matemática* (STEM), e especialmente as mulheres estão sub-representadas. Outro agravante é que a proporção de mulheres que assumem cargos de liderança na indústria ou na academia é ainda menor. Nesse cenário, as instituições de ensino superior têm um papel fundamental no apoio às mulheres e na promoção de um ambiente institucional que busque a igualdade de gênero e o crescimento profissional. Este tópico está, portanto, relacionado a um dos Objetivos de Desenvolvimento Sustentável (ODS[1]). O Objetivo 5 é enquadrado como "Alcançar a igualdade de gênero e empoderar todas as mulheres e meninas ".

Para aumentar a representatividade das mulheres em STEM programas e iniciativas vêm sendo criados. No cenário brasileiro, há o Programa Meninas Digitais - PMD[2] chancelado pela Sociedade Brasileira de Computação (SBC[3]). Por meio de diversos projetos parceiros em instituições de ensino, que são referências no Brasil, e atuando nos pilares do ensino, pesquisa e/ou extensão, disseminam a computação para alunas e professoras ao redor do país (Maciel; Bim; Figueiredo, 2018).

No entanto, para maior efetividade nas iniciativas e para a criação de políticas públicas que atuem nesta direção é necessário primordialmente que se conheçam os fatores que influenciam a falta de equidade de gênero em áreas STEM e que esses sejam disseminados com confiabilidade e visibilidade. Muitas vezes os órgãos, sejam eles públicos ou privados, em contextos maiores ou menores, possuem motivação empírica para criação de ações afirmativas para atrair, manter e desenvolver mulheres que desejam iniciar a carreira em STEM, porém a falta de dados que mostrem os caminhos possíveis de intervenção prejudica ou diminui a adesão por tais ações.

Além disso, embora algumas iniciativas e projetos tenham essa característica de coleta de dados, raramente são publicados de forma aberta e estruturada. Pesquisadores geralmente coletam e analisam dados, mas geralmente quando publicam artigos, o que não garante que os dados estejam facilmente disponíveis para uso público e reutilização, o que tem se mostrado uma demanda crescente (Nunes et al., 2020). Portanto, os dados relacionados a STEM geralmente não são estruturados e não tem boa documentação, em especial, no contexto da América Latina. Outro desafio é que cada pesquisa tem um contexto e nível de análise diferente: projeto, universidade, comunidade e país. Cada contexto pode apresentar significados diferentes para os dados coletados, o que pode prejudicar o entendimento de uso por terceiros.

Em busca de soluções para questões como esta surgiu a rede de pesquisa internacional ELLAS - *Latin American Open Data for gender equality policies focusing on leadership in STEM,* que executa um projeto "Latin American Open Data for gender

[1] https://brasil.un.org/pt-br/sdgs
[2] https://www.sbc.org.br/2-uncategorised/461-Meninas-Digitais
[3] https://www.sbc.org.br/

equality policies focusing on leadership in STEM" proposto e aceito[4] em um edital[5] do *International Development Research Centre* - IDRC. O propósito deste é a criação de uma plataforma de dados abertos contendo dados relacionados à presença de mulheres em STEM, útil para formulação de pesquisas e políticas públicas neste campo (Maciel et al., 2023). Os estudos visam coletar e fornecer de forma aberta e conectada dados sobre mulheres em STEM tais como sobre liderança, fatores que impactam na carreira, iniciativas e políticas públicas. A proposta também inclui a integração dos dados por meio de ontologias e, assim, criar grafos de conhecimento sobre o tema. Este tipo de estrutura possibilita uma representação mais homogênea e comparável entre os dados no âmbito de países da América Latina, a exemplo de plataformas como a descrita por Hyvönen (2020). Assim, concentrando e estruturando esses dados o desenvolvimento de novas políticas e de novas iniciativas se torna mais eficiente pois podem partir de problemas evidenciados pelos dados. A proposta do projeto se concentra em três países vizinhos da América do Sul (Brasil, Bolívia e Peru), com níveis de desenvolvimento bastante diferentes e cujo grupo já vem discutindo sobre esses desafios (Guzman et al., 2020). O projeto é composto de professoras/es, pesquisadoras/es e estudantes com diversas formações, desde Computação, Economia, Psicologia, Pedagogia, Matemática, Física, Química e Administração, todos com interesses em pesquisas sobre mulheres em STEM.

Assim, o presente artigo tem por objetivo apresentar as estratégias adotadas para o desenvolvimento desta plataforma de dados abertos no campo STEM, de forma exploratória descritiva (GIL, 2018), discutindo algumas soluções metodológicas e tecnológicas adotadas para viabilização deste projeto.

2. Contexto do estudo

A utilização de dados estatísticos é essencial e, no Brasil, algumas bases são disponibilizadas de forma aberta e com certa regularidade. Um dos exemplos provém da SBC que disponibiliza anualmente um relatório com os dados dos cursos de computação das universidades brasileiras, cuja fonte de coleta dos dados é o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP[6]). Os dados do INEP contemplam vários cursos, sendo esta compilação feita pela SBC um relatório com enfoque na computação e apesar de ser bastante útil, ele é disponibilizado de forma não estruturada (formato PDF), dificultando análises.

No contexto da América Latina podemos citar também a coleta de dados no Peru e Bolívia, onde a situação não é diferente. No Peru, o Ministério da Educação (MINEDU) disponibiliza um relatório de forma não estruturada, ou seja, em PDF, e com espaços de tempo entre versões que são superiores a um ano (veja por exemplo Torres-Manrique et al., 2021). A fonte dos dados vem do *Sistema de Recolección de Información para Educación Superior* (SIRIES), que conta com informação atualizada, não aberta e que disponibiza dados via solicitação do/a interessado/a. As informações do SIRIES incluem campos tais como quantidade de matrículas por universidade

[4] https://ellas.ufmt.br/ellaswp/manual/
[5] https://idrc-crdi.ca/en/news/gender-stem-research-initiative-announcement-projects
[6] https://www.gov.br/inep/pt-br

pública e/ou privada, gênero, área de conhecimento, ano, entre outros (MINEDU-DIGESU, 2023).

Já na Bolívia, os dados estatísticos referentes ao sistema universitário boliviano são disponibilizados pelo Comitê Executivo da Universidade Boliviana[7]. Este é o órgão central que coordena e programa os objetivos e funções do Sistema Universitário Boliviano, representando-o por meio de ações de planejamento, organização, execução e gestão, reconhecidos no Estatuto Orgânico e Regulamento Geral da Universidade Boliviana Sistema. Apesar de os dados estatísticos serem disponibilizados, nem sempre estão disponíveis no seu portal, no momento da escrita deste artigo. Exemplo disso é o trabalho desenvolvido por Branisa et al. (2021), no qual os autores precisaram coletar dados por meio de entrevistas e grupos focais para explorar o problema de baixa quantidade de mulheres em Tecnologia da Informação na Bolívia.

Em uma abrangência menor, é possível mencionar trabalhos como o de Nunes et al. (2020) que evidenciam a necessidade de dados abertos sobre temas que são base para a compreensão da real situação de egressos em computação. Os autores, que buscavam evidências sobre o mercado de trabalho e os egressos desses cursos na Amazônia, precisaram criar questionários próprios e analisar os dados de maneira "isolada", ou seja, sem fazer parte de uma iniciativa mais estruturada. Outro exemplo de iniciativa que demonstra essa necessidade é o trabalho de Mello et al. (2021) que coletou de forma isolada dados das ações dos projetos, da região sul, parceiros do Programa Meninas Digitais. Esse tipo de dados seria valioso obter de forma estruturada e organizada para não ter apenas os dados de uma região. De forma semelhante, Gindri et al. (2021) e Pereira et al. (2022) necessitaram coletar dados provenientes das redes sociais, da lista de projetos do PMD e/ou de websites.

Essas situações ilustram o desafio de coleta, tratamento e armazenamento estruturado de dados sobre mulheres em STEM. Por mais que existam iniciativas, é necessário que este seja um movimento estruturado e organizado para não criar silos de informação que não ajudam em uma visão mais abrangente sobre o problema e não possam apoiar a tomada de decisões neste campo. Todavia, essa é uma tarefa complexa e que exige esforços teóricos, metodológicos e tecnológicos.

3. Conceitos envolvidos na construção da plataforma

Considerando a intenção deste projeto, o mesmo utilizará uma infraestrutura baseada em Web Semântica[8], sendo os principais conceitos utilizados os de ontologia e grafo de conhecimento semântico, além de os dados serem disponibilizados de forma aberta. Esses conceitos serão explicados nas subseções seguintes.

3.1. Dados abertos

Segundo a definição da *Open Definition*[9], dados abertos são "dados de livre acesso, uso, modificação e compartilhamento para quaisquer propósitos". Também existem 3 normas fundamentais para um dado ser considerado dado aberto: (i) eles devem ser possíveis de

[7] https://ceub.edu.bo/
[8] https://www.w3.org/standards/semanticweb/
[9] https://opendefinition.org/

obter através da Internet em uma forma conveniente e modificável, (ii) eles devem ser distribuídos sob termos que permitam a reutilização livre e (iii) todos podem utilizar sem nenhuma discriminação (Isotani; Bittencourt. 2015). No Brasil, a Lei Nº 12.527 1 , de 18 de Novembro de 2011, trata da abertura dos dados públicos para acesso e utilização de qualquer pessoa física, que traz mais transparência aos dados de interesse da sociedade.

Os dados abertos também podem ser estruturados de uma forma que permita-os serem conectáveis. A partir do momento que eles forem conectados, passam a ser denominados dados abertos conectados. Existe um conjunto de boas práticas para a publicação e conexão de dados, essas são essenciais para que os dados sejam utilizados, de forma automática, por agentes de software (Isotani; Bittencourt, 2015).

3.2 Grafos de conhecimento semântico

Os grafos de conhecimento semântico são estruturas baseadas no formato RDF/OWL (*Resource Description Framework/Web Ontology Language*), que são baseados em triplas, uma forma de representar relacionamentos entre dados (como pode ser visto na Figura 1). Os grafos de conhecimento semântico não são apenas estruturas de armazenamento, como um banco de dados, eles são criados considerando a semântica dos dados utilizando ontologias para descrevê-los e adicionar lógica ao grafo. A lógica é necessária para a criação de novos conhecimentos através de motores de inferência (Fensel et al., 2020).

A ontologia é um campo de estudo da filosofia, que abrange a natureza do ser, da existência e da própria realidade. Para a computação, a definição de ontologia é a representação formal de conceitos de um domínio (Noy e McGuinness, 2001). As ontologias são compostas de classes, propriedades e restrições. As classes são responsáveis por descrever o conjunto de objetos do domínio, as propriedades são as relações entre classes e dados, e as restrições são responsáveis pela parte lógica onde, através de um motor de inferência, é possível gerar novos dados. Por exemplo, na Figura 1 as classes são “Alice” e “Bob” e elas possuem uma relação pela propriedade “is a friend of” (é amigo de), formando uma tripla.

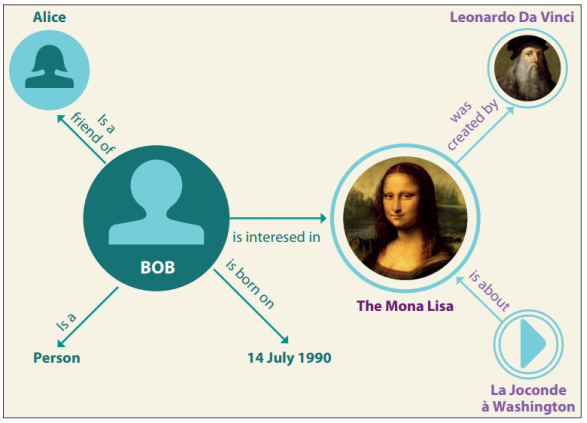

Figura 1. Exemplo de um grafo em RDF (Isotani e Bitencourt, 2015)

3.2.1 Engenharia de ontologias

O desenvolvimento de ontologias é um processo iterativo baseado em 2 passos: (i) criação da ontologia (ii) revisão e refinamento. Esse processo preferencialmente deve ser acompanhado de especialistas do domínio escolhido para se obter uma descrição mais realista e completa do conhecimento em torno do domínio (Noy e McGuinness, 2001). Os passos descritos seguem o processo de desenvolvimento "Ontology 101" de Noy e McGuinness (2001).

Para o processo de modelagem é necessário definir as classes e propriedades que irão constituir a ontologia, para isso começa-se com uma lista com os termos mais importantes do domínio, termos estes que devem representar conceitos no domínio. Neste trabalho de modelagem, dois perfis são essenciais para o sucesso da ontologia: a/o ontologista e a/o especialista de domínio. É responsabilidade da/o ontologista transformar esses termos em classes e propriedades que ditam o relacionamento das classes. A responsabilidade da/o especialista é a criação da lista de termos e a revisão das classes e propriedades definidas bem como definir quais perguntas do domínio o modelo deve ser capaz de responder (chamadas de questões de competência).

Com todas as classes e propriedades criadas é necessária a instanciação dos dados, em outras palavras, a adição dos dados na ontologia. O resultado desse passo é o grafo de conhecimento semântico, que contém a ontologia e os dados instanciados. Para o processamento dos dados presentes no grafo de conhecimento semântico é necessário o armazenamento deste em uma ferramenta criada para trabalhar com dados estruturados no formato de triplas, uma Triplestore. Elas não são apenas responsáveis por armazenar mas também pelo acesso a esses dados, e para isso é utilizada a linguagem SPARQL (SPARQL Protocol and RDF Query Language), uma linguagem de programação que utiliza o conceito de triplas para realizar a busca dos dados.

4. Estratégias para desenvolvimento da plataforma

O projeto tem o prazo de execução de 3 anos (2022, 2023 e 2024), sendo o foco do primeiro e segundo anos a obtenção dos dados e a criação da plataforma, além de atividades como workshops e seminários para fomentar discussões sobre os conceitos que devem estar presentes na plataforma (Maciel et al., 2023). No terceiro ano, a plataforma será promovida por um web site, e aplicações, que utilizam os dados, serão desenvolvidas para ilustrar os potenciais usos dos dados com visualizações em gráficos, por exemplo. A Figura 2 ilustra a estrutura das principais etapas de desenvolvimento da plataforma cujos detalhes são expostos nas seções a seguir.

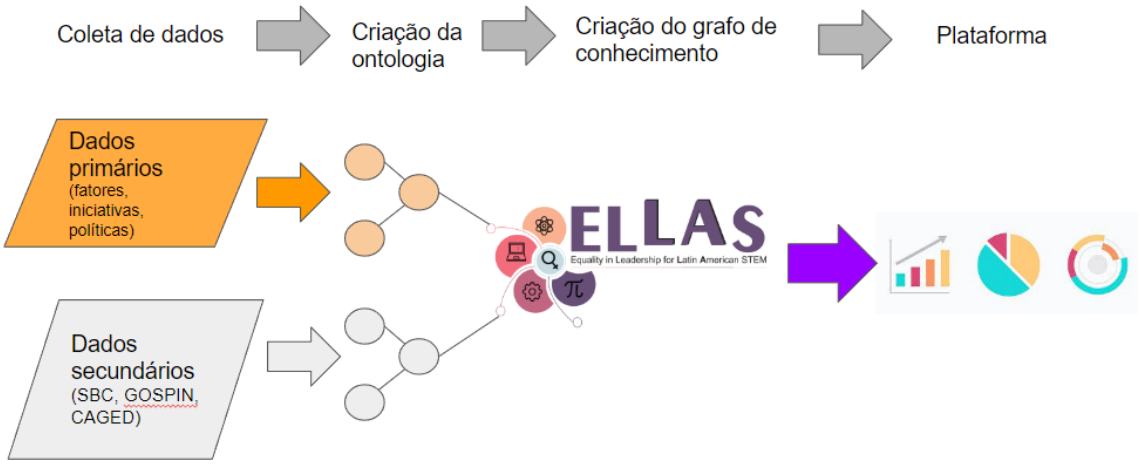

Figura 2. Visão geral da construção da plataforma (autoria própria)

A seguir, essa seção está dividida em 5 subseções em que são dispostos mais detalhes nas etapas de desenvolvimento da plataforma: coleta de dados, criação da ontologia, criação do grafo de conhecimento semântico, desenvolvimento da plataforma e manutenção dos dados. Cada etapa da metodologia de desenvolvimento possui equipes definidas de acordo com a *expertise* de cada pesquisador e estudante. A definição das equipes está intimamente relacionada com o assunto tratado, e com certa diversidade possível de gênero, área de atuação e país.

4.1. Coleta de dados

Como mostra a Figura 2, os dados foram organizados em primários e secundários. Os primários dizem respeito a dados coletados pelos times do próprio projeto, obtidos diretamente de artigos da revisão sistemática de literatura (RSL). Já os secundários são coletados através de portais de dados abertos, como o INEP/SBC, GOSPIN[10] (que é um observatório com dados sobre ciência, tecnologia promovido pela UNESCO), do CAGED (Cadastro Geral de Empregados - com os respectivos salários) e de pesquisas no estilo *survey* desenvolvida pelo grupo de pesquisadoras/es e em universidades e empresas no contexto da América Latina, consideradas alianças parceiras do projeto.

A equipe responsável pela coleta dos dados primários é composta por pesquisadoras/es que trabalham com o tema de equidade de gênero em STEM nos diferentes países que compõem o projeto. Esses são responsáveis por obter dados sobre 3 tópicos relacionados à presença de mulheres em áreas STEM: **fatores** que impactam na escolha e permanência pela área, **iniciativas** desenvolvidas na forma de projetos e **políticas** que já existem com enfoque em mitigar a baixa representatividade feminina nessas áreas. O método utilizado para a coleta dos dados primários foi o de revisão sistemática onde a equipe não só analisou artigos científicos, mas também a literatura cinza. O objetivo é que por meio de RSLs dados que estão descritos nos artigos sejam extraídos e estruturados em planilhas, que mais tarde são usados para modelar a ontologia de integração.

[10] https://gospin.unesco.org/frontend/home/index.php

A equipe responsável por estruturar os dados sobre os **fatores** busca mapear os fatores que influenciam o desenvolvimento da carreira das mulheres em STEM, bem como documentar e analisar iniciativas de sucesso e menos sucesso. É necessário considerar os fatores que influenciam positiva e negativamente as escolhas de carreira das alunas em STEM (Ribeiro, 2020).

A equipe responsável por estruturar dados de **iniciativas** é responsável por mapear projetos e iniciativas que visam a melhoria do cenário atual da representatividade de mulheres em STEM, documentando e analisando as iniciativas de sucesso e menos sucesso. E, por fim, a equipe responsável por estruturar dados de **políticas** existentes é responsável por mapear as políticas e intervenções encarregadas de reduzir a lacuna de gênero.

Todos os dados primários são inseridos em planilhas para estruturá-los minimamente e assim, ser fonte para a construção da ontologia e do grafo de conhecimento semântico. Além disso, a equipe responsável pela criação da ontologia é encarregada de validar as descrições das colunas de dados e a qualidade dos dados.

Para a obtenção dos dados secundários foram criadas 2 subequipes, uma para coletar os dados de portais de dados abertos existentes e uma responsável por criar um questionário estilo *survey* e aplicá-lo para coletar dados que as RSLs não possuem. A equipe responsável pelo *survey* desenvolveu um questionário com base na literatura, e de acordo com boas práticas. O objetivo da aplicação deste *survey* é coletar novos dados não presentes na literatura analisada e, assim, incrementar dados sobre o cenário de liderança feminina em STEM. Vale ressaltar que sua aplicação será feita por empresas contratadas pelo projeto via Fudação de Apoio, com foco de aplicação em universidades e no mercado de trabalho de mulheres em STEM, no Brasil, Peru e Bolívia. Apesar do foco do projeto ser em mulheres, o survey irá coletar dados também de outros gêneros, de maneira a subsidiar análises comparativas.

Ainda, uma tarefa adicional tem sido empreendida, a atualização dos dados coletados na etapa dos dados primários e a geração de outros dados correlacionados que possam ser úteis à plataforma.

4.2. Criação da ontologia

Para a criação da ontologia, são contemplados passos metodológicos da Engenharia de Ontologias supracitada. Para a modelagem e construção da ontologia, uma equipe de especialistas é responsável por analisar as planilhas geradas pela coleta dos dados primários e criar uma ontologia para cada uma delas. Para cada um dos 3 tipos de dados primários há um processo de criação, validação e modelagem da ontologia juntamente com as equipes que coletaram os dados (especialistas do domínio). Caso um problema seja encontrado, a ontologia deverá ser alterada e o processo de revisão será feito novamente. Esse ciclo continua até os resultados serem satisfatórios.

Após a criação das 3 ontologias, será feita uma busca por vocabulários para a descrição das classes e propriedades presentes nas ontologias, por ser uma boa prática a reutilização de termos do domínio. Por fim, a integração de todas as ontologias com os respectivos dados é o que gerará o grafo de conhecimento sobre as mulheres em STEM. Para validar se as ontologias estão representando corretamente os dados, serão utilizadas

as questões de competência que foram definidas pelos especialistas de domínio, que são questões relevantes e que espera-se que a plataforma ajude a responder. Exemplos de algumas questões de competência já definidas são:

- Quais são os fatores que impactam positivamente na inclusão de mulheres na STEM?
- Quais iniciativas estão sendo desenvolvidas para meninas do ensino médio no Brasil ?
- Quais políticas públicas tiveram um bom impacto no Peru?

Como o processo de criação de uma ontologia é cíclico, essa será alterada de acordo com a necessidade proveniente das novas fontes de dados, sejam eles primários ou secundários.

4.3. Criação do grafo de conhecimento semântico

Após a ontologia do projeto ter sido criada, e da coleta dos dados primários e secundários, o grafo de conhecimento semântico será criado.

O desenvolvimento do grafo se dá quando os dados das planilhas "populam" o modelo ontológico criado. Como resultado, um arquivo no formato RDF, contendo todas as triplas, será gerado e este será inserido em uma Triplestore para o consumo dos dados. Esse mapeamento não será estático, ele será utilizado para a criação de novos grafos de acordo com as atualizações dos dados.

A Triplestore será o banco de dados responsável pelo armazenamento dos dados e consumo, através de consultas SPARQL. Esse será parte do back-end do servidor do projeto, no qual qualquer pessoa poderá acessar essas informações através do website do projeto. A ideia é que pesquisas sejam feitas para que os usuários não necessitem de conhecimento em SPARQL para consultar os dados.

4.4. A Plataforma

Com um back-end baseado em ontologias, a plataforma disponibilizará, de forma intuitiva, os dados integrados de projetos, iniciativas e ações realizados em alguns países da América Latina, além de dados estatísticos provenientes de bases abertas como a base do INEP e de pesquisas como a própria survey realizada pelo projeto. Esta integração facilitará os trabalhos de pesquisa de gênero, uma vez que com os dados concentrados o trabalho de busca por informações se torna mais fácil.

Uma das preocupações levantadas pela equipe do projeto foi a utilização de ontologias, e como isso traria uma complexidade na utilização da plataforma, visto que é necessário conhecimento prévio do assunto. Para contornar o problema, o layout do site e as funcionalidades serão desenvolvidos de forma simples, a fim de facilitar a utilização da plataforma para usuários sem conhecimento prévio sobre ontologias. A expectativa é que sejam pesquisadas e desenvolvidas técnicas de interação humano-computador e humano-dados de forma que o usuário não necessite ter conhecimento específico para SPARQL. Para isso, pesquisadoras/es da área de Interação Humano-Computador fazem parte do time do projeto da plataforma.

Para que outras aplicações utilizem os dados da plataforma como input de aplicações de visualização, será disponibilizado o SPARQL endpoint para a Triplestore, pelo qual será possível a conexão com outros grafos de conhecimento semântico.

4.5. Manutenção dos dados

A plataforma do projeto, além de disponibilizar acesso à informação, irá disponibilizar a função de adicionar novos dados primários, quando um novo projeto, iniciativa ou política, por exemplo, for criado. Deste modo, a plataforma se manterá atualizada e relevante para a utilização em novas pesquisas de gênero em STEM. Para alcançar esse objetivo, o mapeamento semântico criado no processo de criação do grafo de conhecimento semântico, juntamente de um script, deverá realizar as funções de forma automatizada, para transformar os novos dados primários em um grafo, e assim ser inserido na Triplestore do projeto. Também, haverá um controle de qualidade para os novos dados seguindo técnicas como as desenvolvidas por Bertucini et al. (2023).

Para os dados secundários, principalmente as bases de dados abertos, devido a atualização frequente dos dados, um pipeline será desenvolvido para a automação da atualização desses dados. Esse pipeline deverá obter as bases assim que elas forem atualizadas, tratar os dados para atender ao escopo do projeto, transformá-los, a partir do mapeamento semântico, para o formato RDF e, por fim, inserir o grafo na Triplestore.

5. Considerações Finais

Os países da América Latina e Caribe sofrem com a sub-representação de mulheres nas áreas de STEM, o que impacta diretamente no desenvolvimento de novas tecnologias. Para reduzir esse problema, iniciativas, políticas e ações são criadas com certa frequência, visando o engajamento de mulheres em STEM, porém a distribuição das informações sobre elas não é concentrada. Isso dificulta o trabalho de análise, que visa a criação de estratégias mais efetivas. Além disso, a distribuição de dados abertos relacionados à presença de mulheres em cursos ou cargos, especialmente em posições de liderança, também sofrem do problema de descentralização, além de muitas vezes não estarem em uma forma estruturada, que facilite a reutilização.

A rede de pesquisa ELLAS e seu projeto foram criados com o intuito de mitigar o problema de descentralização e desestruturação de dados sobre fatores, iniciativas, políticas e ações, e fornecer dados estatísticos sobre a presença de mulheres na área de STEM do Brasil, Peru e Bolívia. O projeto ainda está em desenvolvimento e pretende criar uma plataforma baseada em Web Semântica para facilitar a distribuição e conexão entre esses dados. A plataforma disponibilizará um grafo de conhecimento semântico na qual os dados serão descritos e concentrados para facilitar pesquisas na área.

Neste artigo, apresentamos sumarizadamente a estratégia adotada para criação da plataforma, socializando este conhecimento, possibilitando trocas entre interessados e aprimoramentos à futura ferramenta a ser disponibilizada, que poderá beneficiar toda sociedade que almeja equidade de gênero em STEM.

Agradecimentos

Os autores agradecem ao *International Development Research Centre* - IDRC pela oportunidade e suporte ao projeto, ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq pelas bolsas concedidas e à Fundação UNISELVA pelo gerenciamento administrativo-financeiro do projeto. Em especial, agradecem a toda equipe envolvida no projeto, que atua em diferentes frentes para o sucesso deste e lutam por mais mulheres em STEM.

Referências

Bertucini, O., Berardi, R., Belizario, M., e Kozievitch, N. Garantindo a Qualidade de Dados na Fusão de Dados Conectados: Um caso de uso de SHACL em dados abertos de Mobilidade e Educação de Curitiba. Escola Regional de Banco de Dados, Palmas - PR, 2023

Branisa, Boris; Cabero, Patricia; and Guzman, Indira, "The main factors explaining IT Career Choices of Female Students in Bolivia" (2021). AMCIS 2021 Proceedings. 30.

Brasil. Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Disponível em:<https://presrepublica.jusbrasil.com.br/legislacao/1029987/lei-12527-11>. Acesso em: 04 abr. 2023.

Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., ... and Wahler, A. (2020). Introduction: what is a knowledge graph?. Knowledge graphs: Methodology, tools and selected use cases, 1-10.

Gil, A. C. (2008). "Métodos e técnicas de pesquisa social". São Paulo: Atlas.

Gindri, L., Araújo-de-Oliveira, P., Melo, A. M., Maciel, A., Vargas, K. D. A. R., Otokovieski, M. B. e dos Anjos, R. (2021, July). Mulheres na Computação: de Norte a Sul-Uma Ação de Extensão na Pandemia na Busca pela Integração das Diferentes Regiões do Brasil. In Anais do XV Women in Information Technology (pp. 101-110). SBC.

Guzman, I.; Berardi, R.; Maciel, C.; Cabero Tapia, P.; Marin-Raventos, G.; Rodriguez, N.; Rodriguez, M. (2020). Gender Gap in IT in Latin America. [Disponível em https://aisel.aisnet.org/amcis2020/panels/panels/4/

Hyvönen, E. (2020). Linked open data infrastructure for digital humanities in Finland. DHN 2020.

Isotani, S., and Bittencourt, I. I. (2015). Dados abertos conectados: em busca da web do conhecimento. Novatec Editora.

Maciel, C., Bim, S. A. and Figueiredo, K. S. (2018). "Digital Girls Program: Disseminating Computer Science to Girls in Brazil". In: 40th International Conference on Software Engineering, GE@ICSE018, Gothenburg, Sweden.

Maciel, C., Guzman, I., Berardi, R., Caballero, B. B., Rodriguez, N., Frigo, L., Salgado, L., Jimenez, E., Bim, S. A. and Tapia, P. C. (2023) “Open Data Platform to Promote Gender Equality Policies in STEM”, In Proceedings of the Western Decision Sciences Institute (WDSI). April 2023. Portland Oregon, EUA.

de Mello, A. V., Finger, A. F., Gindri, L., and Melo, A. M. (2021). Mapeamento das ações realizadas pelos projetos parceiros do programa meninas digitais na regiao sul. In Anais do XV Women in Information Technology (pp. 91-100). SBC.

Noy, Natalya F., and Deborah L. McGuinness. "Ontology development 101: A guide to creating your first ontology." (2001).

Nunes, L. H. C., Reis, J. R., Paxiúba, C. M., Ponte, M. J., Nascimento, M. W., and Nascimento, R. P. (2020). Perfil dos Egressos de Computação do Interior da Amazônia no Mercado de Trabalho. In Anais do XIV Women in Information Technology (pp. 254-258). SBC.

Pereira, L. R. R., Souza e Silva, K., Nunes, E. P. S., Maciel, C. (2022). “Perfis em Mídia Social para Meninas e Mulheres com interesse na área STEM e STEAM”. In: Women In Information Technology (WIT), 16, Niterói. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, p. 227-232. DOI: https://doi.org/10.5753/wit.2022.223162.

Ribeiro, K. da S. F. M. (2020). Gênero, Carreira e Formação: O Desenvolvimento da Carreira das Estudantes do Ensino Médio Integrado em Informática. 2020. Tese (Doutorado em Educação). Instituto de Educação, Universidade Federal de Mato Grosso, Mato Grosso.

Torres Manrique, D. S., Pérez Portocarrero, A. J., Carrasco García, F. D., Navarro Véliz, A. N., Obeso Manrique, J. A., Canes Acosta, J. M., ... and Miñan Sánchez, L. F. (2021). Encuesta Nacional de Estudiantes de Educación Superior Universitaria 2019: principales resultados.