A Escassez de Dados Abertos Estruturados em Países Latino-Americanos com Enfoque de Gênero na Educação Superior

Nicole Denes Hildebrand[1], Bruna Oenning Amador[1], Cristiano Maciel[2], Rita Cristina Galarraga Berardi[1]

[1] Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba – PR – Brasil

[2] Universidade Federal de Mato Grosso (UFMT)
Cuiabá – MT – Brasil

{nicolehildebrand, brunaamador}@alunos.utfpr.edu.br, cristiano.maciel@ufmt.br, ritaberardi@utfpr.edu.br

Abstract. *This article analyzes the lack of structured open data in Latin American countries with a gender focus in STEM areas, through research on open data platforms on Higher Education. To this end, the challenges of obtaining this data were discussed, emphasizing the effort required to reuse and integrate the existing data, in order to integrate heterogeneous data and benefit future STEM research more broadly.*

Resumo. *Esse artigo analisa a escassez de dados abertos estruturados em países latino-americanos com enfoque de gênero em áreas STEM, por meio de pesquisas em plataformas de dados abertos sobre Educação Superior. Neste sentido, são discutidos os desafios em obter estes dados, enfatizando o esforço necessário para que os dados existentes possam ser reutilizados e integrados, de modo a poder comparar dados de fontes heterogêneas e beneficiar futuras pesquisas em STEM de forma mais ampla.*

1. Introdução

Apesar da presença feminina na ciência ter aumentado nos últimos anos, nas áreas STEM (Ciência, Tecnologia, Engenharia e Matemática) ainda há uma notável menor presença feminina tanto na academia quanto no mercado de trabalho [Tonini 2019]. Neste sentido, é de suma importância 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 [Berardi et al. 2023]. Assim, mais iniciativas e políticas públicas de fato efetivas podem atuar de forma a mitigar esses fatores e contribuir com o aumento da presença feminina em áreas STEM. Para identificar estes fatores, se faz necessário coletar e analisar diversos dados, entretanto, estes dados geralmente não são disponibilizados em qualidade e formato próprios para reuso, resultando em pesquisas isoladas e que podem apresentar divergências nos fatores encontrados, dificultando o entendimento em um contexto mais amplo, como na América Latina.

Posto isso, o projeto ELLAS, uma parceria entre universidades do Brasil, Bolívia e Peru, financiado pelo *International Development Research Centre - IDRC,* do Canadá, foi criado com o objetivo de contribuir para a geração e o uso de dados abertos e comparáveis entre países, a fim de avaliar políticas e intervenções para reduzir a

lacuna de gênero em STEM; promover a discussão pública com vistas a aumentar o número de mulheres líderes em universidades, indústrias e instituições públicas; e aumentar a conscientização sobre a importância de mulheres em STEM [Maciel et al. 2023]. Para esse fim, é preciso identificar e coletar dados de fontes oficiais existentes para então integrá-los e disponibilizá-los em uma plataforma de dados abertos [Berardi et al. 2023]. No contexto do projeto, a coleta de dados tem sido organizada em dois tipos de dados: "***dados primários***", que compreendem principalmente dados não estruturados em formatos PDF (ou seja, artigos acadêmicos, relatórios etc), dados de mídias sociais e dados coletados por meio de pesquisas (como questionários e entrevistas, p.e.); e "***dados secundários***", que compreendem dados semiestruturados provenientes de sites de organizações nacionais e internacionais, em geral, localizados por meio de portais de dados abertos pré-existentes. Tendo isso em vista, o presente artigo tem como objetivo analisar a escassez de dados abertos estruturados em países latino-americanos parceiros do projeto com enfoque em gênero nos cursos superiores em STEM. Quanto a metodologia, a pesquisa é aplicada com abordagem qualitativa, por meio da elaboração de diagnósticos, identificação de problemas e busca de soluções [Thiollent 2009], com base em observação na Web [Lowe e Pressman 2009].

2. Por que falar sobre Dados Abertos Conectados?

Diversos conceitos foram desenvolvidos nos últimos anos como uma alternativa para a melhoria da qualidade de dados e, recentemente, foram estabelecidos os Dados Abertos Conectados, facilitando o compartilhamento de conhecimento na Web por meio da redução da barreira para a publicação e acesso a dados como parte de um espaço de informação global [Bizer e Berners-Lee 2009 apud Bandeira 2015]. Tendo isso em vista, primeiramente, é preciso definir Dados Abertos e, a partir deste conceito, explicar Dados Abertos Conectados. Segundo *Open Knowledge* (2024), dados são abertos quando qualquer pessoa pode acessar, usar, modificar e compartilhar livremente para qualquer finalidade, e o dado deve estar disponível na internet em um formato compreensível por máquina. Os princípios de dados abertos são: os dados precisam ser completos, primários, atuais, acessíveis, processáveis por máquina, acesso não discriminatório, os formatos não proprietários e livres de licenças [Enap 2016].

Segundo Neves (2013), sobre a Infraestrutura Nacional de Dados Abertos no Brasil, com dados abertos disponíveis, abrem-se possibilidades para a sociedade como a análise mais profunda das informações públicas por meio da correlação de diferentes bases de dados e o desenvolvimento de soluções tecnológicas que fazem uma leitura frequente de bases de dados públicas, para gerar oportunidades de negócio e outros benefícios à sociedade. Deste modo, os Dados Abertos permitem que informações públicas sejam utilizadas livremente na tomada de decisão e produção de conhecimento, de forma que tenham utilidade e possam ser reaproveitados. Com base nisto, tem-se que os Dados Abertos se tornam progressivamente mais poderosos quando conectados [Berners-Lee 2006], posto que possibilita a busca por relações com outros dados de diferentes fontes, de forma a prover contexto e informações úteis. Neste sentido, com Dados Abertos Conectados discute-se o conceito de integrar dados para que a troca de dados e informações possa ser manipulada de forma automática e exista a integração de fontes heterogêneas. Segundo Berners-Lee (2006), trata-se de estabelecer conexões, para que uma pessoa ou máquina possa explorar a teia de dados e, com os dados conectados, encontrar outros dados relacionados.

Posto isso, a utilização de Dados Abertos Conectados abre uma gama de possibilidades para o processamento automático de dados, contribuindo para a descoberta de novos conhecimentos, à integração e integridade de bases de dados, à dificuldade de recuperação e descoberta de dados disponíveis em diversas bases, entre outros [Bandeira 2015]. Estes dados, por sua vez, se tornam úteis para resolver os desafios com relação à qualidade e a confiabilidade de dados sobre mulheres em STEM na América Latina. Para avançar neste sentido, o projeto ELLAS se propõe a gerar e usar dados abertos conectados de países da América Latina. A natureza deste tipo de proposta inclui a integração dos dados por meio de ontologias, uma forma de representação do conhecimento, para então criar grafos de conhecimento sobre o tema, visto que este tipo de estrutura possibilita uma representação mais homogênea e comparável entre dados no âmbito de países da América Latina [Berardi et al. 2023]. Para avaliar a expressividade da ontologia, membros da equipe do projeto (chamados de especialistas do domínio) formam questões que a ontologia deve ser capaz de responder, chamadas de questões de competência, e estas questões justificam a escolha de conceitos e relações na ontologia [Noy e Hafner 1997].

3. Contexto

A reflexão endereçada neste artigo faz parte dos esforços do projeto ELLAS, que organiza a coleta de dados em dados primários e secundários. Os dados primários são dados totalmente desestruturados e normalmente originários de artigos e relatórios em PDF, que então são manualmente analisados e inseridos em planilhas de forma a estruturar os dados encontrados nos textos analisados. Por outro lado, os dados secundários são, em tese, parcialmente estruturados. Neste contexto, atividades foram definidas para cada ano do projeto, e alocadas para pesquisadores/as executarem as tarefas, de forma individual ou em grupo. Neste artigo, o enfoque está na atividade 11, de Recuperação de Dados Secundários [Maciel et al., 2023], que faz parte da segunda fase do projeto e visa coletar dados destas fontes abertas e integrar aos dados primários do projeto, criando grafos de conhecimento conectados por ontologias [Berardi et al. 2023]. Para ser um processo de busca auto sustentável ao longo do tempo, relativamente independente da fonte de subsídio do projeto, automatizou-se este processo de extração, transformação e carga dos dados secundários de modo que a cada atualização da fonte de dados os dados do projeto também são atualizados, como apresentado na Figura 1.

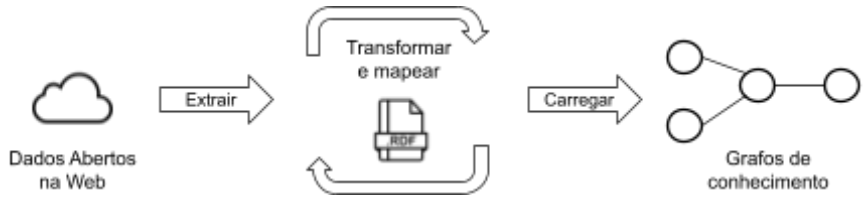

Figura 1. Processo automatizado de ETL (Extrair, Transformar e Carregar)

Para isso, é necessário um trabalho de pesquisa inicialmente manual, realizado por pesquisadores do projeto, filtrando os dados necessários para este contexto, utilizando as ferramentas disponíveis nas plataformas. Em seguida, após geradas planilhas contendo os resultados das buscas, aplica-se o mapeamento para a criação de dados conectados, transformando em triplas RDF. Este formato RDF permite que dados estruturados e semiestruturados sejam misturados, expostos e compartilhados entre diferentes aplicações [W3C 2014], de modo a carregá-los nas ontologias criadas e

conectadas entre si. Assim, será possível fazer *queries* de busca para encontrar informações úteis em diferentes fontes de dados sobre a situação das mulheres nas áreas STEM na América Latina, considerando inicialmente os países parceiros do projeto. A próxima Seção apresenta a aplicação deste processo inicial de busca por fontes de dados abertas estruturadas em enfoque em gênero, subdividido em Brasil, Bolívia e Peru, e finalizando com a busca por dados na base da UNESCO, responsável por "enxergar" a América Latina num contexto mais amplo por ser uma organização mundial da ONU (Organização das Nações Unidas) sobre educação, ciência e cultura.

4. A busca por dados abertos estruturados

Para avançar na busca por dados abertos estruturados, foram escolhidas fontes que incluem dados sobre o Ensino Superior, abrangendo principalmente grandes órgãos do governo ou plataformas com maior relevância no contexto dos países escolhidos (Brasil, Bolívia e Peru), além da análise do portal de dados abertos de cada um deles. Por fim, discute-se o contexto dos dados abertos estruturados disponibilizados pela UNESCO e qual seu papel nesta busca por dados secundários.

4.1. Fontes no Brasil

Para iniciar as buscas por dados brasileiros no contexto da pesquisa, foi tomada como base de pesquisa a Sociedade Brasileira de Computação (SBC), a maior sociedade de Computação da América Latina, tanto em número de sócios quanto em diversidade de iniciativas relacionadas com pesquisa, educação, atuação política e social [Granville e Batista 2020]. A SBC (2024) desenvolve relatórios anuais que disponibilizam um compilado de dados estatísticos da Educação Superior em Computação no País, referentes à distribuição dos cursos, à quantidade de estudantes ingressantes e concluintes e de novos cursos por região, além de estatísticas com enfoque em gênero. No entanto, estas estatísticas são divulgadas somente em formato PDF, o que permite apenas a visualização e/ou a manipulação de forma textual, manual e onerosa dos dados utilizados, não sendo possível reaproveitá-los em novas pesquisas.

Tendo isso em vista, buscou-se a origem dos dados utilizados pelos relatórios da SBC, e foi constatado que eles são uma organização de um recorte para a área de Tecnologia da Informação provenientes dos microdados do Censo da Educação Superior, um levantamento nacional anual de dados estatísticos coordenado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). O INEP oferece informações estatísticas confiáveis, sendo o instrumento de pesquisa mais completo do Brasil sobre as instituições de educação superior. De acordo com INEP (2024), o censo coleta informações sobre a infraestrutura das instituições de educação superior, vagas oferecidas, candidatos, matrículas, ingressantes, concluintes e docentes, nas diferentes formas de organização acadêmica e categoria administrativa. Os dados são fornecidos desde 1995, em formato CSV (valores separados por vírgula), e divididos em dados das Instituições de Ensino e Cursos. No entanto, há uma dificuldade ao fazer recorte de gênero com os dados existentes, visto que o INEP não possui o enfoque que a SBC forneceu sobre os dados, além da falta de integração e a necessidade do uso de um dicionário de dados, que dificultam a análise de forma integrada.

Vale ressaltar que o portal de dados abertos brasileiro [Dados Abertos 2024] tem como pilar a Lei de Acesso à Informação, que tem como objetivo garantir o direito constitucional de solicitar e obter informações dos órgãos e entidades públicas, sendo

que pessoas de qualquer idade e nacionalidade podem pedir informações, além de empresas e organizações [Brasil 2024]. Apesar deste portal ter uma série de dados exclusivos de algumas universidades, os dados do Censo da educação superior do INEP são apresentados logo no início da busca por "Educação superior".

4.2. Fontes na Bolívia

Na Bolívia, não foram encontrados dados disponíveis sobre questões de gênero em áreas STEM. Buscou-se então, a princípio, na base de dados do comitê de universidades *Comité Ejecutivo de la Universidad* (CEUB)[1], organização central coordenadora dos programas, objetivos e funções do sistema universitário boliviano e, portanto, é a principal fonte de dados estatísticos das universidades bolivianas [CEUB 2024]. Mesmo sendo um órgão centralizador de informações a respeito das universidades do país, os dados estatísticos não são todos acessíveis ao público de forma aberta. Para as pesquisas do projeto, tentou-se procurar na sua base de dados estatísticas sobre gênero e número de pessoas matriculadas, total de homens e mulheres por área de conhecimento e, não havia na ferramenta provida pela plataforma nenhuma forma de utilizar filtros de gênero. Há apenas filtros por universidade, área de conhecimento, grau acadêmico, matrículas novas, sede e gerais e titulados, sendo possível procurar estatísticas por ano a partir de 2007. Tendo isso em vista, o que pode ser feito para encontrar dados de gênero estritamente quantitativos, usando a ferramenta estatística dentro do portal do comitê, é utilizar os filtros disponíveis, seja por universidade, sede, etc. Após selecionar o ano desejado, o site gera um gráfico de barras contendo os números totais e divididos por gênero do que foi buscado, no entanto, é possível fazer download deste gráfico somente como imagem (formato não estruturado). Essas informações não estão disponíveis para serem baixadas em outros formatos que facilite conectá-las a outras fontes de dados.

A segunda grande plataforma de acesso livre acerca da publicação de dados públicos da Bolívia é o *Portal de Datos Abiertos*, o qual propõe-se em ser um portal do Estado Plurinacional da Bolívia de dados abertos, acessíveis, estruturados e reutilizáveis [Datos Abiertos 2024]. A plataforma conta com uma divisão em categorias de economia, cultura, educação, entre outras, porém, a maioria dessas categorias não possui realmente algum dado. E, a respeito de educação, até o presente momento, encontram-se somente dados de enquetes sobre a inclusão digital realizadas em 2019 e um conjunto de dados com descrição, localização e permanências de estabelecimentos educativos, sendo a última atualização em novembro de 2016. Por fim, é importante ressaltar que essa escassez de dados tão evidente de fontes bolivianas, como visto nos sites governamentais apresentados, é possivelmente devido à falta de uma lei nacional de acesso público à informação. De acordo com Villanueva e Taborga (2017), que apresentam o relatório de 2017 da UNESCO, a Bolívia é um dos poucos países da América Latina a não ter nenhuma lei que garante acesso à informações públicas.

4.3. Fontes no Peru

Tendo em vista que não foram encontrados documentos com estatísticas voltadas para mulheres em áreas STEM em sites disponíveis no Peru, buscou-se bases de dados em que o recorte de gênero poderia ser realizado, apesar de não ser um processo trivial. A primeira plataforma encontrada foi TUNI: *Sistemas Información Universitaria*, sendo a

[1] Dado o último acesso em 1 de abril de 2024, a ferramenta de visualização de dados estava fora do ar.

fonte oficial de informação sobre o sistema de ensino superior universitário [TUNI 2024]. A plataforma é gerenciada pela Superintendência Nacional de Ensino Superior Universitário (SUNEDU), que realiza relatórios bienais sobre a realidade Universitária no Peru, buscando consolidar e armazenar informações [SUNEDU 2024]. Desta forma, pretende-se disponibilizar aos cidadãos toda a informação relevante para a tomada de decisão sobre o ensino superior universitário, além de fornecer às organizações de ensino um meio online para a apresentação de informações sobre o ensino superior universitário de maneira oportuna e eficiente. De modo a possibilitar o recorte de gênero na plataforma, foram acessadas as estatísticas da comunidade estudantil.

A plataforma apresenta dados subdivididos em Matriculados, Ingressantes, Candidatos e Dados Históricos. Considerando as três primeiras abas, existem estatísticas visuais e filtros para busca neste conjunto de dados, de acordo com a aba escolhida, podendo ser visualizados e baixados em formato de planilha do Excel. Nota-se que os dados são recentes, a partir de 2020 nas abas Matriculados e Ingressantes e somente 2022 em Candidatos. Nas categorias, em geral, há o total de estudantes por categoria e por universidade, dados sobre tipo de gerenciamento, cursos categorizados em grupos, nome do curso e dados sobre localização da instituição de ensino. Além disso, existe a opção de considerar mais variáveis na tabela, como variáveis de gênero, faixa etária e pessoas com deficiência. Estes dados, apesar de poderem ser encontrados e baixados, estão dispersos entre as categorias e existem falhas na usabilidade, podendo gerar dificuldade e exigir tempo para encontrar o que precisa.

Na aba de Dados históricos é possível acessar dados anteriores aos anos citados, vale ressaltar que a disposição dos elementos na tela não seguem o mesmo padrão das outras categorias e que esta aba também é subdividida em categorias, sendo elas: Por Universidade (2014-2016), Por Programa (2014-2016) e Por Universidade (1955-2013). Na primeira categoria, existem variáveis de gênero nos dados, que são apresentados utilizando filtros, gráficos visuais e uma tabela, que pode ser exportada como Excel ou CSV, facilitando a manipulação destes dados. Considerando a segunda categoria, mantém-se os filtros e uma tabela principal, existindo variáveis de feminino e masculino, que também podem ser baixadas em formato CSV ou Excel. Por fim, na terceira categoria há uma visualização em Power BI (ferramenta da Microsoft de análise de dados que possibilita a criação de *dashboards*), e existem ainda mais subcategorias, entretanto há a possibilidade de baixar estes dados somente em formato Excel. Em resumo, existem dificuldades de usabilidade da plataforma TUNI, para encontrar os dados necessários e os botões de baixar os dados, por exemplo, dado que as informações são descentralizadas. Além disso, os dados são disponibilizados em baixa qualidade quando considerado o formato CSV, tendo em vista a não utilização de um separador, dados não necessários como o valor total.

O governo do Peru também disponibiliza a *Plataforma Nacional de Datos Abiertos* (PNDA), que permite encontrar, explorar e reutilizar dados governamentais de forma simples, segura e confiável [PNDA 2024]. Esta plataforma foi criada no contexto da Lei sobre Governo Digital, que tem o objetivo de estabelecer a estrutura de governança para o governo digital no Estado e o regime jurídico para o uso de tecnologias digitais na Administração Pública [Peru 2024]. Tendo isso em vista, ao analisar a aba de Educação na plataforma, há uma série de dados sobre Universidades e em formato CSV, juntamente com outros dados sobre educação. No entanto, estes dados

são disponibilizados em diferentes *links* e por universidade, muitas vezes relacionados somente a um ano e uma categoria em específico, dificultando a extração dos dados.

4.4. UNESCO

A Organização das Nações Unidas para a Educação (UNESCO) tem por objetivo principal usar da educação, ciência, cultura, comunicação e informação para fomentar respeito e entendimento mútuo [UNESCO 2024]. Neste sentido, possui duas plataformas próprias que disponibiliza, tanto para o público geral, quanto para cientistas, políticos e empresários, dados sobre seus próprios projetos e outros tipos de projetos, iniciativas, programas de iniciativa pública e privada de diversos países e que não envolve necessariamente a participação da UNESCO. Ambos os domínios buscam ter um formato mais acessível de trazer esses tipos de dados sobre o desenvolvimento da ciência e tecnologia ao redor do mundo, já que é disponível para todos, deve ser fácil de navegar e intuitivo. A UNESCO, por meio destas plataformas, também propõe-se a manter dados livres e abertos sobre o maior número de países possível, com a proposta de incluir países subdesenvolvidos e emergentes, não só focados no norte global.

Tendo isso em vista, o domínio *Global Observatory of Science, Technology and Innovation Policy Instruments* (GO-SPIN) é uma plataforma de livre acesso que oferece aos seus usuários bancos de dados com gráficos e ferramentas analíticas para ajudar os tomadores de decisão. Focada em mostrar dados de diversas formas sobre ciência, tecnologia e inovação, contendo políticas, estatísticas, dentre diversos outros dados relacionados à ciência e educação global. Sendo assim, a plataforma contém dados e relatórios sobre diversos países e suas respectivas políticas educativas e de ciências e, inclusive, sobre inclusão de minorias nas áreas STI (ciência, tecnologia e inovação), relatórios socioeconômicos, recomendações sobre open science, etc. O GO-SPIN também possui uma ferramenta de busca por “Women in STEM”, permitindo filtrar os resultados por países selecionados. Em sua essência, o que se encontra nessa seção do site são iniciativas e/ou premiações sobre políticas para promoção de equidade de gênero nas áreas STEM [GOSPIN 2024]. Contudo, esses dados não vêm estruturados e sim, em forma de relatórios possíveis de baixar em formato PDF, que não permite uma manipulação de dados de forma automatizada. Com relação aos países da América Latina, o GOSPIN possui poucos dados sobre suas políticas para mulheres nas ciências e, em muito do que foi encontrado, há problemas com os dados, como: Dados muito antigos e não atualizados; Duplicação de dados, como no Programa Mulher e Ciência, que consta com uma versão em espanhol de 2017 e uma em inglês de 2020 e não há dados sobre o Brasil em português brasileiro, o que dificulta o acesso por parte de brasileiros sobre informações de seu país de residência; Links nos relatórios que não existem ou não funcionam, impossibilitando o acesso à origem do dado.

Por outro lado, o Core Data Portal tem seu foco no comprometimento da UNESCO de transparência, compartilhando dados e relatórios de seus trabalhos e parcerias ao redor do mundo [Core Data Portal 2022]. A plataforma, também de acesso ao público, provê informações essenciais e relatórios dos trabalhos feitos por eles ao redor do mundo. Os dados são, em sua essência, sobre os programas e orçamentos da Unesco e as principais estratégias da organização e os resultados de seus projetos. Vale ressaltar que a plataforma é recente, tendo dados somente a partir de 2022 e é atualizada trimestralmente. Posto isso, um dos problemas encontrados ao utilizar a plataforma é que para encontrar as iniciativas e projetos que a UNESCO contribui em relação às

questões de equidade de gênero em STEM, é preciso primeiro entender os programas principais da UNESCO e suas divisões de resultados para enfim encontrar dados de interesse. O *Major Programme* que foi utilizado nessa pesquisa é o de Ciências Naturais o qual, segundo UNESCO (2022), tem como prioridade contribuir para a prioridade global da Organização de promover equidade de gênero, e prestar particular atenção aos países menos desenvolvidos. Os resultados desse programa estão divididos em *outcomes* e *outputs*, sendo que os *outcomes* são baseados nas prioridades de ação da UNESCO que se traduzem em cada um dos objetivos estratégicos da nova *Medium -Term Strategy* para 2022-2029. *Outputs*, por sua vez, são *outcomes* dentro de cada uma das áreas de seus programas e mensurados com indicativos de quantidade e qualidade.

Após entender as áreas dos projetos e seus resultados com base nas métricas da UNESCO, foi possível obter alguns dados em relação às áreas da STI e STEM de países da América Latina. Porém, ainda que o Core Data tenha dados melhores estruturados e possíveis de serem baixados em formato CSV, há o mesmo problema que não se tem tantos dados disponíveis a respeito desses países. Isso ocorre principalmente devido a maioria dos projetos da própria UNESCO serem focados na África, já que faz parte de uma de suas prioridades responder às necessidades dela. Além disso, alguns países da América Latina, mesmo não sendo considerados ricos ou desenvolvidos, tecnicamente não são considerados "Países menos desenvolvidos" pelas métricas da UNESCO, não tendo enfoque em seus projetos tanto quanto outros países considerados mais pobres.

5. Desafios em obter dados estruturados

No âmbito da ciência aberta, existe uma tendência mundial para dar acesso livre aos periódicos científicos e essa demanda se estende agora para o acesso livre e inteligível dos dados gerados pela pesquisa científica [Sayão e Sales 2014]. No entanto, a atual oferta de dados na web ainda tem ocorrido em formatos que impõem limitações quanto à reutilização, dado que a maioria são consumidos apenas por humanos, não permitindo que sejam reutilizados de forma automatizada por agentes de software [Wood et al. 2013 *apud* Bandeira 2015]. Sendo assim, ainda que o crescimento contínuo da quantidade de dados produzidos pelos diversos segmentos da sociedade confere a esses recursos a condição de componente fundamental para a pesquisa científica moderna [Sayão e Sales 2014], para que estes dados possam ser reutilizados de forma a agregar valor em pesquisas futuras, é preciso que sejam estruturados e de qualidade, de modo a facilitar a sua compreensão e manipulação.

Neste mesmo sentido, o desafio se estende dado que no contexto da América Latina, há uma carência de pesquisas e base de dados que coletem e divulguem abertamente dados de qualidade sobre a atual realidade das mulheres dentro das áreas de STEM nesses países [García-Holgado et al. 2019], o que dificulta ainda mais o entendimento do problema e a criação de iniciativas, políticas e ações para reduzi-lo. Trazendo para o contexto do projeto ELLAS, foi percebido que a **distribuição das informações** sobre mulheres em STEM não é concentrada, existindo dados esparsos em diferentes pesquisas e correspondentes de diferentes bases de dados, que nem sempre estão divulgadas para manipulação e reutilização, resultando na **perda de dados** que poderiam gerar grandes contribuições em pesquisas futuras. Além disso, em uma mesma plataforma, muitas vezes **não são disponibilizados grandes volumes de dados em um único acesso**, sendo necessário buscar dados muitas vezes equivalentes, mas distribuídos em diferentes acessos, como em relatórios periódicos.

Outro desafio existente, é que ainda que estes dados existam e sejam divulgados, ainda se percebe uma limitação nos dados encontrados, muitas vezes utilizam-se **formatos** que inviabilizam o reuso dos dados e que necessitará um grande retrabalho para novas análises. Neste sentido, a manipulação destes dados é um processo não trivial e custoso para ser realizado de forma manual, além de limitar a **reutilização destes dados de forma automatizada**, conforme foi proposto para os dados secundários. Um exemplo presente nos casos apresentados foi a utilização do formato PDF, que embora seja um formato relevante para um público específico, é um formato proprietário não estruturado que impõe barreiras de uso por agentes que consomem de forma automatizada [Alcantara et al. 2015], inviabilizando novas análises e reuso.

Além disso, existe a **dificuldade em encontrar o que deseja** nas plataformas também foi presenciada, posto que algumas plataformas não são de fácil acesso e compreensão, com baixa usabilidade especialmente para novos usuários, dificultando a busca por informações. A **forma como as informações são apresentadas**, por sua vez, também dificulta ainda mais a usabilidade, existindo muitos materiais desatualizados, descentralizados e links quebrados durante a navegação. Portanto, a coleta de dados secundários não é trivial e há muito no que avançar no quesito de disponibilização e qualidade dos dados, especialmente quando o enfoque se dá para a problematização da sub-representação de mulheres em STEM. Assim, quanto maior a capacidade dos sistemas de informação de oferecer dados de pesquisas livremente, de forma que possam ser interpretados e reutilizados pelo maior número possível de pesquisadores de diversas áreas, maior será o grau de transparência, de reprodutibilidade e de eficiência do processo de geração de conhecimento científico, e maior será a amplitude de aplicação dos projetos de pesquisa para a sociedade [Sayão e Sales 2014].

Para ilustrar os desafios mencionados neste artigo, desenvolvemos a Tabela 1 que evidencia a escassez de dados e seus impactos das dificuldades supramencionadas. A ideia da tabela é registrar os resultados parciais do processo executado pela equipe do projeto para coletar e integrar dados educacionais de mulheres em STEM nesses países. Para isso, foram selecionadas algumas das questões formuladas pela equipe do projeto ELLAS que deverão ser respondidas com o auxílio da plataforma e, ao tentar respondê-las, detalhes puderam ser percebidos, conforme a Tabela 1. A saber, BR referencia Brasil, BO referencia Bolívia e PE referencia Peru.

É possível perceber que a busca por respostas nestas plataformas não é trivial e que os dados não contêm um padrão na forma que são disponibilizados e organizados, dificultando a comparação entre estes dados para se explorar uma visão no contexto da América Latina. Além disso, pode-se perceber que para perguntas de pesquisa mais simples e que não envolvem muitos dados, o comportamento das plataformas se mantém parecido e é possível encontrar as respostas, no entanto para perguntas complexas e que envolvem mais de uma variável, são encontrados diversos desafios com relação aos dados disponibilizados nas plataformas, como a falta de dados e a quantidade de arquivos necessários para encontrar o resultado desejado. Também, a necessidade de navegação por muitos *links* dificulta o acesso a estes dados, sendo um processo custoso e lento. Outro ponto é que as plataformas se limitam a uma única língua utilizada, dificultando o entendimento dos dados em outros idiomas, e algumas exigem a necessidade de estudar um vocabulário, por exemplo ao utilizar um dicionário de dados, diminuindo ainda mais a semântica dos dados disponibilizados.

Por fim, os desafios mais apresentados foram: a escassez de dados nestas plataformas, que supostamente deveriam conter estas informações, e a limitação do ano em que estes dados estão disponíveis, além dos dados coletados não trazerem o mesmo nível de detalhamento com relação aos cursos e às universidades. Isto evidencia a importância de existirem leis que contribuam para a distribuição de informação de forma aberta e transparente, devendo ser levado em consideração o quão recentes são estas leis no contexto latino americano, sendo inexistentes em países como a Bolívia, o que corroboram ainda mais para a precariedade destes dados.

Tabela 1. Questões de Competência por Plataforma

Questões de competência	Plataforma		Arquivos baixados	Formato	Nº de links	Estudar vocabulário	Há filtros?	Língua	Desafios	Há resposta?
Quantas mulheres ingressaram em cursos de computação em 2020?	BR	SBC	1	PDF	2	Não	-	PT	-	Sim
		Inep	1	CSV	1	Sim	Não	PT	Manipulação dos dados e recorte de gênero	Sim
	BO	CEUB	-	-	-	-	-	-	-	Não
		Datos abiertos	-	-	-	-	-	-	-	Não
	PE	TUNI	1	Excel	1	Não	Sim	ES	-	Sim
		PNDA	Vários	CSV	Vários	Sim	Sim	ES	Pode não conter dados de 2020	Não
Quantas mulheres concluíram bacharelados em cursos de STEM em universidades públicas de 2018 a 2022?	BR	SBC	1	PDF	2	Não	-	PT	Dados até 2021; Não há tipo de universidade e curso; Somente computação	Não
		Inep	5	CSV	5	Sim	Não	PT	Recorte de gênero	Sim
	BO	CEUB	-	Gráfico de barras	2	Não	Não de gênero	ES	Não há como baixar como planilha; Não há filtro de gênero; Não é possível separar por período ou tipo de faculdade.	Não
		Datos abiertos	-	-	-	-	-	-	Não foram encontrados dados	Não
	PE	TUNI	-	Excel	-	Não	Sim	ES	Não há tipo de curso; Não há concluintes, apenas anterior a 2017	Não
		PNDA	Vários	CSV	Vários	Sim	Sim	ES	Não contém todos os anos	Não

6. Considerações finais

Os dados e conjuntos de dados de pesquisas providenciam as evidências necessárias para conferir veracidade, autenticidade e capacidade de reprodutibilidade ao corpo de conhecimento publicado nos periódicos, o que parece ser fundamental para o progresso científico [Sayão e Sales 2014]. Entretanto, o presente artigo discute que os dados sobre mulheres em áreas STEM no contexto da América Latina muitas vezes são inexistentes ou apresentam uma série de desafios para seu (r)euso, sendo evidenciada a necessidade de dados de qualidade e que sejam disponibilizados de forma estruturada e aberta. Dito isso, considerando o contexto de dados secundários do projeto ELLAS, as pesquisas em STEM também serão beneficiadas com a disponibilização destes dados conectados, posto que possibilitará a comparação entre as diferentes bases de dados, auxiliando a evidenciar o problema, produzir novos conhecimentos e responder novas questões de pesquisas. Sendo assim, estes dados são essenciais, entre outros, para o entendimento dos fatores que influenciam a sub-representatividade das mulheres nestas áreas e para auxiliar nas tomadas de decisão para alterar esta realidade. Como trabalhos futuros, busca-se expandir a análise para outros fatores que se relacionem com este contexto e para outros países latino-americanos, ampliando o entendimento do problema.

Agradecimentos

Gostaríamos de agradecer ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e ao *International Development Research Centre* do Canadá - IDRC, bem como a todas as pessoas pesquisadoras envolvidas no projeto ELLAS.

Referências

Alcantara, W., Bandeira, J., Barbosa, A., Lima, A., Ávila, T., Bittercourt, I., & Isotani, S. (2015). Desafios no uso de Dados Abertos Conectados na Educação Brasileira. In Anais do IV Workshop de Desafios da Computação aplicada à Educação, (pp. 11-20). Porto Alegre: SBC. doi:10.5753/desafie.2015.10036

Bandeira, J., Ávila, T., Alcantara, W., Barbosa, A., Bittencourt, I. I., & Isotani, S. (2015). Dados abertos conectados para a educação. In Anais. Porto Alegre, RS: SBC. Recuperado de http://www.br-ie.org/pub/index.php/pie/article/view/3551/2937

Berardi, R., Auceli, P., Maciel, C., Davila, G., Guzman, I., & Mendes, L. (2023). ELLAS: Uma plataforma de dados abertos com foco em lideranças femininas em STEM no contexto da América Latina. In Anais do XVII Women in Information Technology, (pp. 124-135). Porto Alegre: SBC. doi:10.5753/wit.2023.230764

Berners-Lee, T. (2006) Linked Data. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>.

Brasil (2024) LAI para Cidadãos - Conheça seus direitos. Disponível em: <https://www.gov.br/acessoainformacao/pt-br/assuntos/conheca-seu-direito>.

CEUB (2024). Disponível em: <https://ceub.edu.br>.

Dados Abertos (2024). Disponível em < https://dados.gov.br/home>.

Datos Abiertos (2024). Disponível em: <https://datos.gob.bo/>.

Enap (2016). Elaboração de Plano de Dados Abertos. Diretoria de Comunicação e Pesquisa SAIS - Área 2-A - 70610-900 — Brasília, DF. Disponível em <https://repositorio.enap.gov.br/bitstream/1/3152/1/M%C3%B3dulo%201%20-%20Conceitos%20de%20Dados%20Abertos.pdf>.

García-Holgado, A., Camacho Díaz, A., & García-Peñalvo, F. J. (2019). Engaging women into STEM in Latin America: W-STEM project In M. Á. Conde-González, F. J. Rodríguez-Sedano, C. Fernández-Llamas, & F. J. García-Peñalvo (Eds.), TEEM'19 Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality (Leon, Spain, October 16th-18th, 2019) (pp. 232-239). New York, NY, USA: ACM. doi10.1145/3362789.3362902

Granville, L. Z.; Batista, T. V. (2020) O papel das sociedades científicas. Computação e sociedade: a profissão, Cuiabá-MT: EdUFMT Digital, v. 1, 1ª edição, p. 177 - 193.

Inep (2024). Censo da Educação Superior. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior>.

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.

Neves, O. M. D. C. (2013). Evolução das políticas de Governo aberto no Brasil. Anais do VI Congresso Brasileiro de Gestão Pública – CONSAD. Brasília, Brasil.

Noy, N. F.; Hafner, C. D. (1997) The state of the art in ontology design: A survey and comparative review. AI Magazine, vol. 18, p. 53–74.

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

Peru (2024). Ley de Gobierno Digital. Disponível em: <https://www.datosabiertos.gob.pe/ley-de-gobierno-digital>.

PNDA (2024). Disponível em: <https://www.datosabiertos.gob.pe/>.

Sayão, Luis Fernando; Sales, Luana Farias. (2014). Dados abertos de pesquisa: ampliando o conceito de acesso livre. RECIIS - Revista Eletrônica de Comunicação, Informação e Inovação em Saúde, Rio de Janeiro, v. 8, n. 2, p. 76-92, jun.

SBC (2024). Educação Superior em Computação - Estatísticas. Disponível em: <https://www.sbc.org.br/documentos-da-sbc/category/133-estatisticas>.

SUNEDU (2024). Superintendencia Nacional de Educación Superior Universitaria - Informes Bienales. Disponível em: <https://www.gob.pe/institucion/sunedu/colecciones/5716-informes-bienales>.

Thiollent, M. (2009). Metodologia de Pesquisa-ação. São Paulo: Saraiva.

Tonini, A. M.; Araújo, M. T. de. (2019) A participação das mulheres nas áreas de STEM (Science, Technology, Engineering and Mathematics). Revista de Ensino de Engenharia, v. 38, n. 3, p. 118-125. Disponível em: <http://revista.educacao.ws/revista/index.php/abenge/article/view/1693>. Acesso em: 25 agosto 2021.

TUNI (2024). Disponível em: <https://www.tuni.pe/>.

UNESCO (2022). Approved programme and budget 2022-2025: first biennium 2022-2023. UNESCO, 7, place de Fontenoy, 75352 PARIS 07 SP. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000380868>.

UNESCO (2024) Disponível em: < https://www.unesco.org/en>.

GO-SPIN (2024) Disponível em: < https://gospin.unesco.org/ >.

CORE DATA PORTAL (2024) Disponível em: < https://core.unesco.org/en/home >.

Villanueva, E. R. T.; Taborga, S. V. (2017). Assessment of Media Development in Bolivia based on UNESCO's Media Development Indicators - United Nations Educational, Scientific and Cultural Organization, 7, place de Fontenoy, 75352 Paris 07 SP, France and UNESCO Quito/ Representation for Bolivia, Colombia, Ecuador and Venezuela, p. 11-35. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000246906 >.

W3C (2014). Resource Description Framework (RDF). Disponível em: <https://www.w3.org/RDF/ >.