UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

VANESSA LAZARIN DE SOUZA

ELABORAÇÃO DE UMA ONTOLOGIA QUE ENDEREÇA A PRESENÇA DE MULHERES EM CURSOS DE COMPUTAÇÃO NO BRASIL

CURITIBA
2023

VANESSA LAZARIN DE SOUZA

ELABORAÇÃO DE UMA ONTOLOGIA QUE ENDEREÇA A PRESENÇA DE MULHERES EM CURSOS DE COMPUTAÇÃO NO BRASIL

Elaboration of an ontology to address women's presence on computer courses in Brazil

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para a obtenção do título de Bacharel em Sistemas de Informação do Curso de Sistemas de Informação da Universidade Tecnológica Federal do Paraná.

Orientadora: Dr[a] Rita Cristina Galarraga Berardi
DAINF - Departamento Acadêmico de Informática - UTFPR

CURITIBA
2023

VANESSA LAZARIN DE SOUZA

ELABORAÇÃO DE UMA ONTOLOGIA QUE ENDEREÇA A PRESENÇA DE MULHERES EM CURSOS DE COMPUTAÇÃO NO BRASIL

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para a obtenção do título de Bacharel em Sistemas de Informação do Curso de Sistemas de Informação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 20 de junho de 2023

Drª Rita Cristina Galarraga Berardi
Universidade Tecnológica Federal do Paraná

Dr. Cristiano Maciel
Universidade Federal do Mato Grosso

Drª Nádia Puchalski Kozievitch
Universidade Tecnológica Federal do Paraná

CURITIBA
2023

Para as mulheres que vieram antes de mim e para as que ainda hão de vir.

AGRADECIMENTOS

Às minha avós, Carolina e Adalziza. À minha mãe, Izabete. À minha madrinha, Maria Lúcia. Às minhas tias Dumicilia, Cecília, Claudia, Riusa, Paola, Vera, Dalva e Edna. A vocês - grandes mulheres da minha vida - meu mais profundo amor.

Ao meu pai, Sidiney e ao meu irmão Gustavo. Às minhas amigas & amores - em especial à Emanueli, Hayssa, Hell e Luana pela presença, cuidado e paciência. À minha terapeura, Penélope, por me ajudar a voltar integralmente para mim.

À minha orientadora Rita, pelo espaço seguro para a construção desse trabalho e por me lembrar que há muitos horizontes possíveis. Ao Tacla e meus colegas de PETECO, por me ajudarem, lá no início, a entender o alcance social da computação.

Aos meus colegas de pesquisa Bruna e Pedro. Às muitas mulheres incríveis - colegas e professoras - com quem compartilhei e compartilho minha jornada acadêmica. Vocês me inspiram e me fazem sentir pertencente.

À mim mesma, por não desistir.

"Technology is not neutral. We're inside of what we make, and it's inside of us. We're living in a world of connections — and it matters which ones get made and unmade."

Donna Haraway

RESUMO

SOUZA, Vanessa L.. Elaboração de uma ontologia que endereça a presença de mulheres em cursos de computação no Brasil. 2023. 60 f. – , Universidade Tecnológica Federal do Paraná. Curitiba, 2023.

A presença de mulheres nas áreas de STEM *(Science, Technology, Engineering and Math)* é uma questão que vem ganhando cada vez mais relevância. Contudo ela esbarra na falta de dados para a construção de análises consistentes. O *Equality for Leadership in Latin American STEM* (ELLAS) - do qual participam Bolívia, Peru e Brasil - tem como objetivo elaborar uma plataforma de *Linked Open Data* que ajude a preencher essa lacuna. Dentro de tal empreendimento temos a criação de ontologias e a reestruturação para RDF de dados já existentes como pontos centrais na criação de LOD. O presente trabalho se localiza dentro do ELLAS e colabora com (1) a triplificação (RDF) de dados do Censo de Educação Superior do Inep, (2) a criação de uma metodologia para elaboração de ontologias, para ser usada dentro do projeto e que possibilite análises sobre a presença e permanência de mulheres nas áreas de STEM e (3) a instanciação de tal ontologia no contexto do Ensino Superior na área da computação no Brasil.

Keywords: Ontologias. Gênero. Computação. STEM. *Linked Open Data*.

ABSTRACT

SOUZA, Vanessa L.. Elaboration of an ontology to address women's presence on computer courses in Brazil. 2023. 60 f. – , Universidade Tecnológica Federal do Paraná. Curitiba, 2023.

Women's presence in STEM (Science, Technology, Engineering and Math) areas is an issue that has been gaining relevance. However it collides with the lack of data for the construction of consistent analyses. The Equality for Leadership in Latin American STEM (ELLAS) project - in which Bolivia, Peru and Brazil participate - aims to develop an Linked Open Data (LOD) platform to help fill that gap. Within such entrepreneurship, ontology creation and the restructuring (RDF) of previously existing data are key elements on the LOD creation. The present work is located inside ELLAS and collaborates with (1) the triplication (RDF) of data from the Higher Education Census from Inep, (2) the creation of a methodology for elaborating ontologies - to be used within the project and that enables analysis about the presence and permanence of women in STEM areas and (3) its instantiation in the context of Higher Education in the field of computing in Brazil.

Keywords: Ontologies. Gender. Computer. STEM. Linked Open Data.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CSV	*Comma-Separated Value*
DAG	Dados Abertos Governamentais
HTML	*HyperText Markup Language*
IDRC	*International Development Research Centre*
KG	*Knowledge Graphs*
LOD	*Linked Open Data*
OWL	*Ontology Web Language*
QC	Questão de Competência
RDF	*Resource Description Framework*
RDFS	*Resource Description Framework Schema*
SBC	Sociedade Brasileira de Computação
SPARQL	*SPARQL Protocol and RDF Query Language*
STEM	*Science, Technology, Engineering and Math*
STI	*Science, Technology and Innovation*
TIC	Tecnologia de Informação e Comunicação
URI	*Universal Resource Identifier*
USDA	*United States Department of Agriculture*
XML	*Extensible Markup Language*
W3C	*World Wide Web Consortium*

LISTA DE CONSULTAS

SUMÁRIO

1 INTRODUÇÃO

A sub-representação de mulheres na computação é um fato concreto e uma discussão que tem ganhando espaço a nível mundial nos últimos anos. Iniciativas que fomentam o debate ao redor dessa questão vem surgido em diversos espaços: nos debates sociais e políticos, no mundo empresarial, nas universidades. No Brasil, iniciativas como PrograMaria[1] e {reprograma}[2] procuram empoderar meninas e mulheres a adentrar a computação através da disponibilização de cursos e conteúdo técnico, organização de conferências e encontros entre mulheres da área. O Departamento de Informática da UTFPR campus Curitiba conta com as iniciativas Emíli@s[3] e TIchers[4] - participantes do Programa Meninas Digitais[5], chancelado pela Sociedade Brasileira de Computação (SBC) - que realizam diversas ações junto à comunidade interna e externa desde 2013. A baixa presença de mulheres em cursos de computação é uma questão mundial e o Brasil está incluso nessas estatísticas, com a participação de mulheres nos cursos de computação abaixo de 16% (MACIEL; BIM; FIGUEIREDO, 2018).

A presença de mulheres em espaços de poder é - para além do domínio da computação - outro debate que vem se ampliando. E a computação, como espaço de poder que é, privilegia àqueles mais perto da visão cartesiana do que as ciências denominam de "sujeito neutro". Este sendo a persona de homem racional europeu branco heterossexual burguês. Quanto mais longe deste ideal se está, com mais estranhento será percebida sua presença em tais espaços.

Se a presença de mulheres na computação em geral já é uma problemática, ao galgar patamares superiores em hierarquias e organogramas organizacionais a situação se agrava. Diversos trabalhos apontam as questões históricas atreladas a esse contexto, muitos estudos de caso apontam a atual gravidade da situação, no entanto há uma ausência de informações para que se façam possíveis análises mais embasadas e consistentes.

Dado que vivemos numa sociedade dependente de Tecnolologias de Informação e Comunicação (TIC), em que já em 2010, mais de 70% do PIB de países como Japão, Canadá e Alemanha já derivava de bens intangíveis associados à informação, o acesso à elas é absolutamente determinante para uma compreensão consistente da realidade (ISOTANI; BITTENCOURT, 2015).

E para criar informações confiáveis, acessíveis e utilizáveis por todos precisamos de dados de qualidade, abertos e conectados para que sejam utilizados, reutilizados e redistribuídos livremente. Dados que possam ser lidos por pessoas e processados por

[1] <https://www.programaria.org>
[2] <https://reprograma.com.br>
[3] <https://emilias.dainf.ct.utfpr.edu.br>
[4] <https://tichers.ct.utfpr.edu.br>
[5] <https://meninas.sbc.org.br>

máquinas, que possam se conectar com outras fontes, fomentando a geração de novos dados baseada no consumo dos que existiam antes, se movimentando num ciclo contínuo de geração de informação (Figura 1).

Figura 1 – Ciclo de Vida da Informação

Fonte: Isotani e Bittencourt (2015)

O funcionamento e o crescimento da economia de países globais e a gestão de governos e corporações é totalmentente dependende de TIC e do alto volume de dados (*Big Data*) que elas tanto geram quanto consomem. Tal quantidade chega a casa de zettabytes 10^{21}, e seu ciclo (Figura 1) envolve uma temporalidade que nos mostra a importância que a estruturação e a conexão desses dados têm para simplificar e facilitar a recuperação da informação, impulsionando assim a geração de conhecimento.

O interesse público por Dados Abertos Conectados se dá exatamente por seu potencial de criar conhecimento e fomentar uma interpretação mais embasada da complexa realidade da sociedade e suas problemáticas. Desde 1997 a W3C (*World Wide Web Consortium*), consórcio que cuida dos padrões e das tecnologias relacionados ao desenvolvimento da Web, estabelece padrões e orientações para que o potencial máximo da Web seja atingido. Tal potencial é o de participação ativa da Web na solução de problemas humanos e sociais, e o uso de Dados Abertos Conectados é parte central dessa construção.

É nesse ponto - inserido em um projeto de maior magnitude - que se encontra o presente trabalho. O projeto internacional *Latin American open data for gender-equality policies focusing on leadership in STEM* é realizado pela rede de pesquisa *Equality for Leadership in Latin America STEM* (ELLAS)[6] e pretende criar uma base de dados abertos e conectados que sirva de fomente a construção de políticas públicas e privadas que enderecem a presença de mulheres em STEM na América Latina. O projeto é financiado

[6]Por praticidade, ELLAS é o acrônimo usado ao longo do texto para se referir ao contexto do projeto. Os objetivos gerais do projeto podem ser encontrados no Apêndice A.
<https://idrc-crdi.ca/en/news/gender-stem-research-initiative-announcement-projects>

pela instuição de pesquisa canadense *International Development Research Centre*[7] (IDRC), com participação de pessoas pesquisadoras na Bolívia, Peru e Brasil e duração estimada de 36 meses.

Dados de referência e estatísticas no recorte de gênero e STEM são difíceis de encontrar, isolados e muitas vezes exigem conhecimentos de manipulação de dados. Arquivos em CSV, como o do Inep, utilizado na seção 4.3, existem mas caem dentro dessa categoria. Em contrapartida, a facillidade de extração de significado que uma ontologia tem e sua ampla conectividade com outras fontes de informação são os alicerces do seu grande potencial de uso.

Dentro do ELLAS, as ontologias servirão de base para o desenvolvimento de uma aplicação de dados abertos e conectados, acessível a pessoas de dentro e de fora da área acadêmica, com informações sobre os três países do projeto e pronta para incorporar outras fontes. Tal centralização de referências sobre gênero e STEM possibilitará pesquisas mais bem fundamentadas e confiáveis na área, permitindo elaborações materiais - como políticas públicas e privadas - que se baseiem em uma compreensão mais objetiva da realidade.

O presente trabalho colabora nessa concepção com a elaboração de uma metodologia para construção de ontologias dentro do projeto e com a instanciação de uma ontologia para o contexto de mulheres na computação no Brasil.

1.1 Objetivos

O objetivo geral da presente pesquisa é projetar e instanciar uma ontologia que possibilite marcar e criar recortes de gênero ao representar dados referentes a presença e permanência de mulheres nas áreas de computação. Para alcançar tal intento foram estabelecidos os seguintes Objetivos Específicos (OE):

- (OE1) Reestruturar os dados disponibilizados pelo Inep no Censo de Educação Superior de 2021 para responder à questòes pertinentes a presença de mulheres no contexto da computação brasileira
- (OE2) Definir uma metodologia para elaboração de ontologias, a ser utilizada no ELLAS
- (OE3) Instanciar a metodologia criada utilizando os dados do Inep e validá-la

Para alcançar tais objetivos se planeja o uso de conceitos da área de estudos de gênero e engenharia de ontologias, o levantamento de dados sobre mulheres na computação através do Censo do Ensino Superior do Inep e a aplicação de técnicas computacionais do campo de engenharia de ontologias, incluindo o sistema computacional de gerenciamento de conhecimento *Protégé*.

[7] <https://idrc-crdi.ca/en/about-idrc>

O escopo deste trabalho se localiza dentro das atividades da fase 1, que vai de janeiro de 2022 a junho de 2023, do projeto ELLAS. Subsequentemente existirão as fases 2 e 3 - de duração similar.

2 REVISÃO DE LITERATURA

Neste capítulo são apresentados, contextualizados e discutidos conceitos que embasam o desenvolvimento deste trabalho. Embora esta pesquisa tenha um objetivo bastante concreto - o desenvolvimento de uma ontologia - ela só faz sentido dentro de seu contexto social. Assim sendo temos a computação como meio para contribuir com um projeto maior e criar possibilidades de solução para um problema humano e social - a presença e permanência de mulheres nas áreas de STEM. Na primeira seção definimos o que são carreiras de STEM, em 2.2 estabelecemos o que são dados abertos e conectados. Na seção seguinte definimos *Knowledge Graphs* (KG) e ontologias. Por fim definimos termos pouco habituais na computação.

2.1 Carreiras de STEM

O acrônimo STEM foi cunhado na década de 1990 por Judith A. Ramaley, então diretora do *National Science Foundation's Education and Human Resources Division* nos Estados Unidos, para descrever as áreas relacionada à ciência, tecnologia, engenharia e matemática (ENGLISH, 2016). Tais áreas são comumente vistas como pertencentes exclusivamente às ciências exatas, entretanto há debates que consideram agronomia e ciências da saúde como também pertencentes a esse grupo. Koonce et al. (2011) estabele a definição do termo STEM em duas perspectivas: educacional e ocupacional, sendo a primeira mais próxima da pesquisa acadêmica e a última mais ampla, incluindo mais frequentemente engenharia agrônoma e de alimentos.

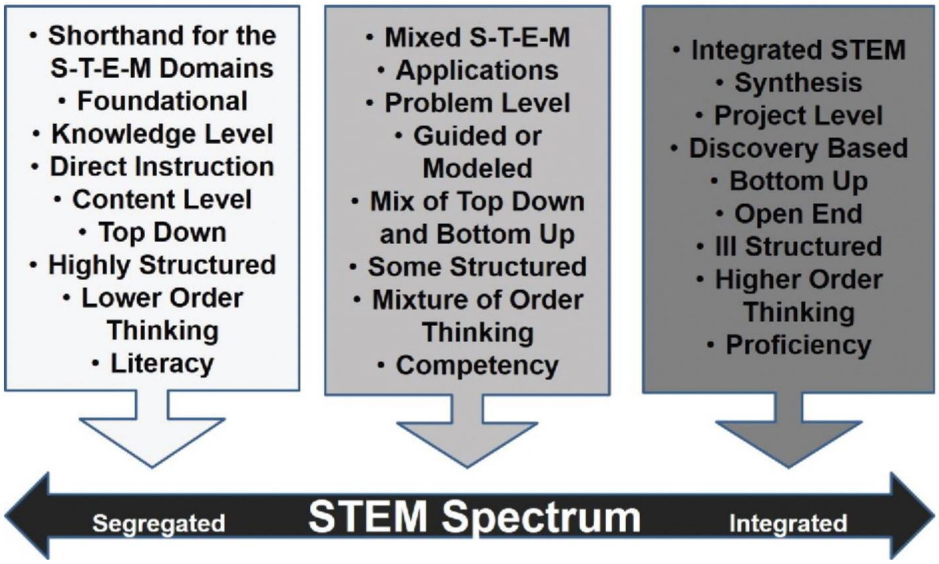
O debate em torno de como definir o que STEM abrange é amplo e aberto a diferentes interpretações. English (2016) explica o conceito de STEM integrada, em contrapartida ao de STEM segregada, como uma área interconectada e interdependente de outras, onde o transpassar de limites se faz necessário para que problemas complexos e transdisciplinares do mundo real sejam analisados.

Na imagem 2, Nadelson e Seifert (2017) analisa as áreas de STEM como um espectro onde em um extremo estão áreas segregadas e no outro áreas integradas. Esse é um debate contínuo e não há um concenso absoluto e uma definição estática do termo "carreiras de STEM". No presente trabalho - para limitação de escopo - partiremos de uma definição educacional e segregada desse conceito, e dentro deste teremos como foco a área da computação.

2.2 Dados abertos e conectados

O potencial da Web de gerar conhecimento reside em suas conexões e depende da forma como tais conexões são feitas e de que artefatos são conectados. A Web de Documen-

Figura 2 – Espectro das áreas de STEM



Fonte: Nadelson e Seifert (2017)

tos, formato atual da rede, estabelece um conjunto de padrões: os URIs (*Uniform Resource Identifier*) como identificadores globais e únicos para cada página; o HTTP ((*Hyper Text Transfer Protocol*) como protocolo de transferência e mecanismo de acesso universal; e o HTML (*HyperText Markup Language*) como formato padrão para representação de conteúdo.

O estabelecimento desses padrões foi responsável pela transformação que a *World Wide Web* causou na forma que compartilhamos informações no mundo, entretanto suas limitações cerceiam o poder de gerar conhecimento da rede. Para que pessoas e máquinas possam extrair mais significado do conteúdo publicado na Web se faz necessário o uso de tecnologias semânticas que utilizem recursos informacionais mais eficientemente.

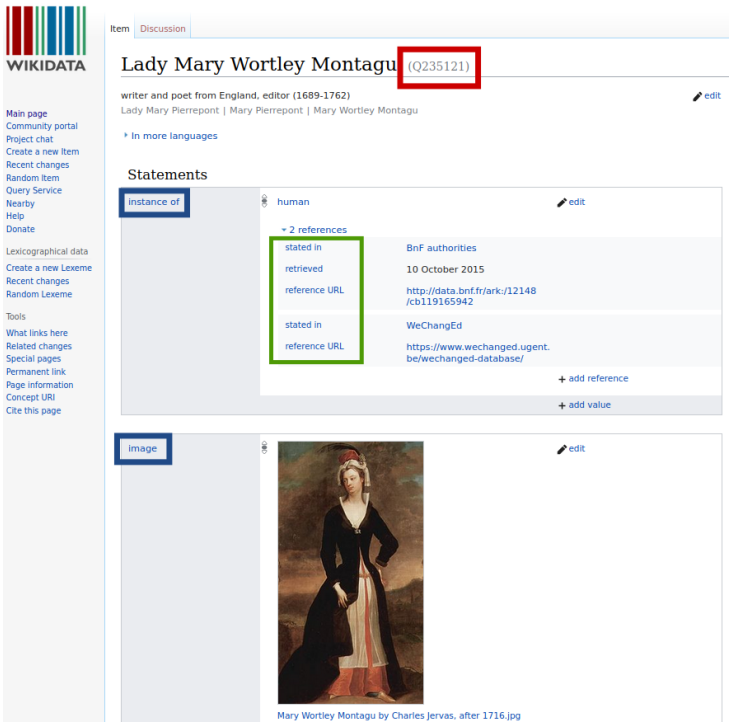
A Web de Dados utiliza metadados antes invisíveis para dar significado às conexões entre dados estruturados e ela também se baseia em padrões: os mesmos que a Web de Documentos para identificação e acesso (URI e HTTP), o padrão RDF (*Resource Description Framework*) para representação de conteúdo e a linguagem SPARQL para o acesso aos dados. Dessa forma temos Dados Conectados que podem ser utilizados mais eficientemente e conectados a outros dados facilmente, maximizando o potencial de trabalho integrado entre pessoas e máquinas.

O conjunto de dados do projeto *WeChangEd* - apresentado no capítulo anterior - segue o modelo de dados empregado pela Wikidata - que consiste de Itens, Propriedades e Identificadores Únicos (THORNTON et al., 2021). No exemplo do item Lady Mary Wortley Montagu (figura 9), temos como identificador na Wikidata Q235121 (retângulo vermelho). As propriedades '*instance of*' e '*instance of*' (retângulos azuis) são predicados para declarações e três propriedades - '*stated in*', '*iretrieved*'e '*reference URL*'(retângulo verde)

- são usadas como predicados e qualificadores e provém referências para as informações apresentadas. Cada item representa um recurso é um sujeito (s), conectado a objetos (o) através de predicados (p), formando triplas (s, p, o).

A criação de triplas (s, p, o) através de um conjunto de declarações de sujeito, predicado e objeto estrutura o modelo RDF de representação de dados. O RDF expressa a relação do sujeito com o objeto, através de uma propriedade que representa a natureza dessa relação, o predicado (ISOTANI; BITTENCOURT, 2015). Complementando a RDF, há a *Resource Description Framework Schema* (RDFS), que possibilita a expansão semântica dos dados através da utilização de classes e criação de categorias, com a possibilidade de hierarquizá-las.

Figura 3 – Captura de tela do item Lady Mary Wortley Montagu na Wikidata



Fonte: Wikidata[1]

Aqui temos triplas que relacionam recursos com objetos que podem ser apenas dados ou outros recursos. No exemplo da Figura 2 temos as triplas:

1. Sujeito <Lady Mary Wortley Montagu> que se relaciona como <instance of> do objeto <human>
2. <Lady Mary Wortley Montagu> como sujeito, <image> como predicado e sua imagem como objeto

A criação de triplas (s, r, o) através de um conjunto de declarações de sujeito, predicado e objeto estrutura o modelo RDF de representação de dados. O RDF expressa a relação do sujeito com o objeto, através de uma propriedade que representa a natureza dessa

relação, o predicado (ISOTANI; BITTENCOURT, 2015). Complementando a RDF, há a *Resource Description Framework Schema* (RDFS), que possibilita a expansão semântica das dos dados através da utilização de classes e criação de categorias, com a possibilidade de hierarquizá-las.

Para que alcancem seu potencial através de relações semânticas, dados digitais estruturados conectados devem ser também abertos. Dados abertos são aqueles que podem ser utilizados, reutilizados e redistribuídos livremente, além de respeitar às normas fundamentais de disponibilidade e acesso, reuso e distribuição e participação universal (ibidem). A riqueza do trabalho desenvolvido por THORNTON et al., ilustrado na Figura 9, reside nas conexões dos dados gerados pelo projeto com outros já existentes, o que não seria possível caso os dados fossem meramente conectados, pois bases de dados fechadas não se comunicam.

O compartilhamento de dados em formato aberto permite o cruzamento com dados de diferentes fontes, para que sejam livremente utilizados e reutilizados pela sociedade de forma a promover a geração e uso de novos dados de alta qualidade baseada em dados anteriormente consumidos (ISOTANI; BITTENCOURT, 2015). Esse ciclo, ilustrado na Figura 1, para acontecer efetivamente e ser capaz de prover informações úteis e gerar novos conhecimentos, precisa de dados padronizados, acessíveis, e conectados. Os chamamos de Dados Abertos Conectados, do inglês *Linked Open Data*.

Dados Abertos Conectados carregam em si "um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na Web, com o intuito de criar uma Web de Dados" (ibidem). É importante citar que nem todos os dados conectados são abertos (e vice-versa) e que redes privadas que se baseiam nesse conceito existem, no entando o enfoque dessa pesquisa é no seu uso público.

Figura 4 – Caneca das 5 Estrelas de Dados Conectados

Fonte: Isotani e Bittencourt (2015)

Dados Abertos Conectados costumam ser classificados pelo "Sistema de 5 Estrelas",

criado por Berners-Lee (2006) e mantido pela W3C, que avalia seu grau de abertura, denotando a facilidade destes dados de serem enriquecidos ao se conectarem com outros dados conforme a Figura 4.

No exemplo do projeto *WeChangEd*, construi-se uma aplicação web semântica para a exibição interativa dos dados disponibilizados na Wikidata, Wikimedia e Wikipedia. Nela podemos acessar galeria de mídia, artigo da Wikipedia, linha do tempo, mapa de lugares e pessoas relevantes, declarações da Wikidata, biblioteca de materiais e outros links relacionados às editoras mapeadas no projeto (figura 10. O desenvolvimento dessa aplicação se fez possível pelos dados utilizados serem Dados Abertos Conectados com classificação 5 Estrelas, significando que os dados utilizam formato RDF e SPARQL, que seus dados se conectam a outros dados de forma a fornecer um contexto sobre eles.

Uma aplicação importante baseada em dados abertos e conectados são os DAG - Dados Abertos Governamentais ou Dados Abertos Públicos. Dentre as motivações centrais dos DAG há o acesso transparente aos gastos de fundos públicos, fatos e informações, proporcionando às pessoas maior compreensão sobre o estado e em suas interações com o governo. Além disso, o uso de DAG possibilita endereçar problemas urbanos e sociais complexos (ARZBERGER et al., 2004). No cenário internacional, o compartilhamento de dados científicos contribui para uma melhor compreensão sobre aspectos de saúde pública, especialmente a área de epidemiologia e biocomplexidade pelo menos desde 2002 (ibidem).

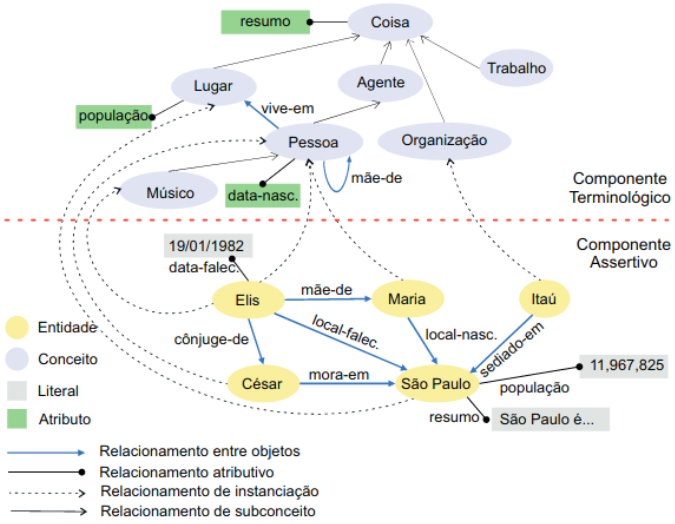
2.3 *Knowledge Graphs* e Ontologias

Grafos de Conhecimento, do inglês *Knowledge Graphs* (KG), são formas de representar conhecimento que proveem uma estrutura semântica adequada para que sistemas computacionais os processem e ao mesmo tempo tem uma representação próxima à linguagem natural, facilitando a interpretação por pessoas. KG são formados por nós, que descrevem objetos, e arestas, que os conectam (SILVA; ZIVIANI; PORTO, 2019). Costumeiramente, nós representamos objetos do mundo real, suas categorias e asserções sobre elas utilizando-se de triplas <sujeito, relação, objeto>.

KG são representações formais de conhecimento e compostos pelos componentes terminológico (ontológico) e assertivo (de entidades) para que o propósito de que máquinas alcancem o significado da informação seja alcançado (Figura 5). Os conjuntos de dados que compõem um KG devem ser estruturados, normalizados e conectados, e o modelo RDF é o padrão estabelecido pela W3C e amplamente utilizado. A Figura 5 ilustra a camada ontológica e o componente assertivo do KG, o primeiro mais abstrato e o segundo instanciando o primeiro.

Ontologia é um termo que vem da filosofia, atrelado ao estudo da natureza, origem, métodos e limites do conhecimento na epistemologia e à natureza da realidade na metafísica. Dentro da computação o termo costuma se referir à um artefato de engenharia constituído por um vocabulário controlado usado para descrever uma determinada realidade juntamente

Figura 5 – Exemplo de *Knowledge Graphs*



Fonte: Silva, Ziviani e Porto (2019)

a um conjunto de assunções explícitas sobre o significado pretendido das palavras desse vocabulário (GUARINO, 1998). Em resumo, uma ontologia é uma forma de representação de conhecimento. E é a ontologia que delimita a estrutura do *Knowledge Graphs*.

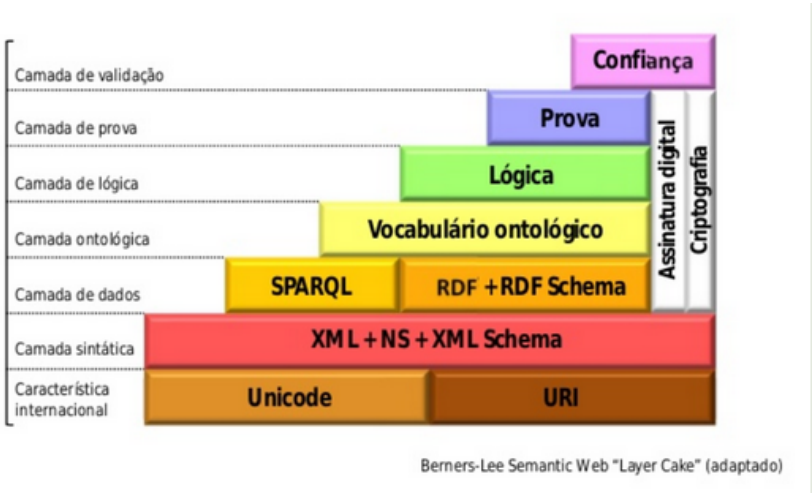
KG e ontologias alcançam seu maior potencial a partir de Dados Abertos Conectados. Seus padrões são elementos centrais na evolução da Web de Documentos para uma Web Semântica ou Web de Dados. Através dessa mudança computadores podem não só ler, mas extrair sentido de dados que encontram na Web. Assim ela também se faz dependente da publicação de dados abertos e de qualidade. E quanto mais conectados estes dados forem, mais sentido os processos de inteligência artificial intrínsecos à Web Semântica extrai deles.

A Figura 6 ilustra a arquitetura de Camadas da Web Semântica em um modelo simplificado desenvolvido por Berners-Lee (2006) chamado de "Bolo de noiva". Ele explicita o papel de padrões de formato (RDF) e consulta de dados (SPARQL) na estrutura da Web Semântica e busca tornar a implementação desta mais factível e fácil de ser visualizada por aquelas que a utilizam e desenvolvem.

2.4 Metodologias para Desenvolvimento de Ontologias

A engenharia de ontologias é um campo de estudo emergente que abrange métodos, metodologias e ferramentas para o desenvolvimento e gerenciamento de ontologias (ABDELGHANY; DARWISH; HEFNI, 2019). Para a escolha de uma metodologia para o presente trabalho, foi realizada a revisão da literatura sobre metodologias para desenvolvimento de ontologias, dentre as quais se destacam:

Figura 6 – Camadas da Web Semântica



Fonte: Berners-Lee (2006)

1. *Ontology development* 101 (NOY; MCGUINNESS, 2001): metodologia clássica e amplamente reconhecida na literatura. Tem uma abordagem *top-down* de sete passos, onde primeiro se determina o domínio e escopo da ontologia e por último são criadas suas instâncias
2. SABiO - *Systematic Approach for Building Ontologies* (FALBO, 2014): composta de 5 fases - cada uma com com tarefas atreladas a papéis determinados - começando com a identificação do propósito e dos requerimentos e terminando no teste da ontologia
3. *NeOn Methodology* (SUÁREZ-FIGUEROA; GÓMEZ-PÉREZ; FERNÁNDEZ-LÓPEZ, 2012): metodologia baseada em cenários, com foco em sua reusabilidade e na conexão entre ontologias distintas, estruturando 9 dos cenários mais prováveis dentro desse espectro
4. AMOD - *Agile Methodology for Ontology Development* (ABDELGHANY; DARWISH; HEFNI, 2019): inspirada em metodologias ágeis de desenvolvimento de software, considera as fases de pré-jogo, desenvolvimento e pós-jogo

Foi realizada uma análise comparativa entre as metodologias escolhidas e seus resultados podem ser vistos na tabela 1. Tendo em vista a escalabilidade pretendida e a natureza colaborativa do projeto ELLAS, foram considerados os seguintes aspectos: (1) colaboratividade, (2) conectividade com outras ontologias, (3) reusabilidade por outras ontologias, (4) etapas para sua construção, (5) elaboração de documentação e (6) ferramentas mencionadas para seu desenvolvimento.

O presente trabalho se embasa a partir da análise de dados e plataformas já existentes, assim sendo a criação de uma ontologia do zero não é realista com a natureza desta pesquisa. Nessa perspectiva, as metodologias NeOn e SABiO 2.0 são as que se

percebem como melhores candidatas.

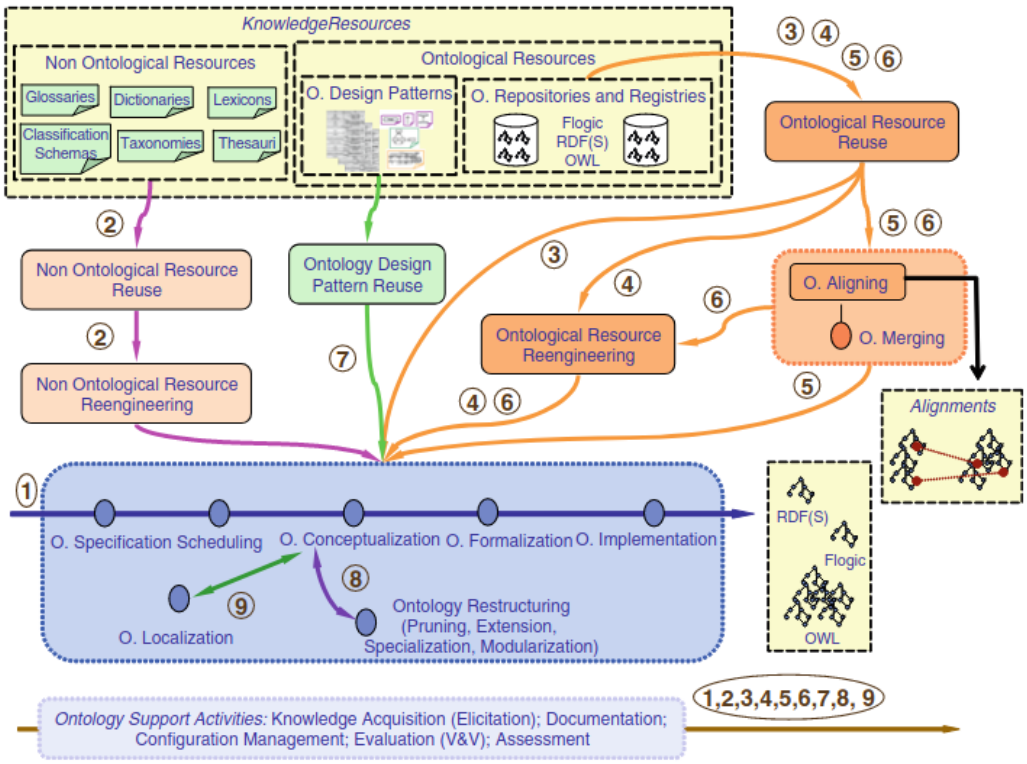
Os cenários apresentados pela NeOn como os mais prováveis para o desenvolvimento de ontologias são: (1) Da especificação à implementação; (2) Reutilização e reengenharia de recursos não ontológicos; (3) Reutilização de recursos ontológicos; (4) Reutilização e reengenharia de recursos ontológicos; (5) Reutilização e fusão de recursos ontológicos; (6) Reutilização, fusão e reengenharia de recursos ontológicos; (7) Reutilização de padrões de projeto de ontologias (ODPs); (8) Reestruturação de recursos ontológicos e (9) Localização de recursos ontológicos (SUÁREZ-FIGUEROA; GÓMEZ-PÉREZ; FERNÁNDEZ-LÓPEZ, 2012). A Figura 7 desenha o caminho que cada um desses cenários percorre ao longo da elaboração da ontologia.

No escopo do projeto ELLAS está a análise de dados já existentes como parte da elaboração da pesquisa, logo, dentro da NeOn, nos encontramos no Cenário 2: Reutilização e re-engenharia de recursos não-ontológicos. Esse cenário se adequa ao ponto de partida do ELLAS pelo fato da aplicação semântica a ser desenvolvida ser específica e aplicada, o que permite uma ontologia que use referências e que seja mais "leve", em contrapartida a ontologias de alto nível de abstração que são mais "pesadas"e buscam representar um universo amplo. Como o objetivo do projeto é se conectar e somar com iniciativas já existentes, a análise de recursos não-ontológicos relacionados é imprescindível e consta nos objetivos 1 e 2 do projeto (Apêndice A), sendo detalhado como objetivo específico (OE1), na seção 1.1 do presente trabalho.

O fluxo do cenário começa com (1) a análise para reuso e reengenharia de recursos não-ontológicos já existentes e então segue para a linearidade de (2) especificação, (3) planejamento, (4) conceitualização, (5) formalização e (6) implementação. A primeira parte inclui a pesquisa, avaliação e seleção das entidades não-ontológicas apropriadas, seguida de processos de engenharia reversa e transformação sobre tais entidades, resultando em um modelo conceitual que é finalmente utilizado como base na elaboração da ontologia. Paralelamente acontecem atividades de documentação, aquisição de conhecimento, avaliação de conteúdo, gerenciamento de configuração e análise de alinhamento com as necessidades das/os usuárias/os.

Considerando-se o contexto do projeto ELLAS, a colaboratividade ao longo de todas as etapas do processo é determinante - afinal somos em dezenas de colaboradores ao redor de diversos países. Para um projeto dessa dimensão, a documentação é um processo cuja priorização é crucial, e a divisão de papéis e a ênfase na documentação são pontos fortes da metodologia SABiO 2.0. Falbo (2014) define os seguintes papéis: (1) especialista de domínio, a pessoa especialista no domínio da ontologia e que fornece o conhecimento a ser modelado e implementado na ontologia de domínio; (2) usuária/o da ontologia, representando o público por quem se intenciona que a ontologia seja usada; (3) engenheira/o de ontologias, responsável pela ontologia de referência, ou seja, pelas fases iniciais do desenvolvimento da ontologia; (4) designer de ontologias, papel que cuida do

Figura 7 – Cenários para a elaboração de ontologias da metodologia NeOn



Fonte: Suárez-Figueroa, Gómez-Pérez e Fernández-López (2012)

design da ontologia operacional; (5) programadora/o de ontologias, pessoa a implementar a ontologia operacional e (6) testadora/o de ontologias. Esses papéis podem ou não pertencer a pessoas distintas, sendo comum que haja intersecção. Engenheira/o e designer de ontologia, por exemplo, são responsabilidades relacionadas e frequentemente atribuídas a mesma pessoa (FALBO, 2014).

2.5 Outras definições

Para além dos termos da área da computação, é importante explicitar a definição de gênero utilizada no presente trabalho. Entendemos como gênero a construção social, cultural e psicológica que determina comportamentos, estereótipos, atitudes e expectativas. Como aponta TANNENBAUM et al. (2019, p. 138):

> "A experiência de gênero inclui três dimensões interrelacionadas. Identidade de gênero diz respeito a como indivíduos e grupos se percebem e apresentam dentro de contextos específicos. Relações de gênero descrevem as dinâmicas de poder entre indivíduos com identidades de gênero distintas. Finalmente, normas de gênero se referem às regras - implícitas e explícitas, na família, no

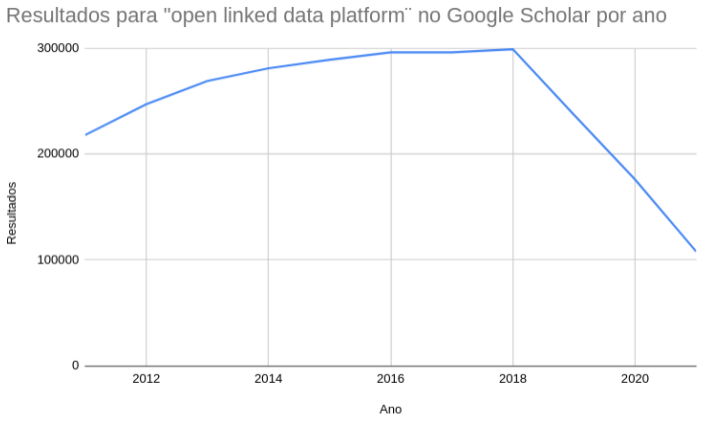
trabalho, nas instituições e culturais - que nos são impostas."

2.6 Trabalhos correlatos

Esta sessão traz a análise de trabalhos recentes correlatos a este. A pesquisa se deu através dos repositórios Google Acadêmicos, IEEE Xplore e Periódicos da Capes com o uso do termo "*open linked data platform*"e filtro de resultados entre os anos de 2017 e 2022. A escolha dos artigos se deu pela quantidade de citações contabilizada e análise de resumos. Na seleção, foram privilegiados trabalhos acadêmicos que descrevem a criação de aplicações web semânticas de interesse público. Inicialmente foram procurados trabalhos dentro do tema que tangencionassem gênero, contudo os resultados se mostraram bastante escassos - fato que confirma a importância de projetos como o ELLAS e pesquisas como a do presente trabalho. Dentro da classificação de metodologia de pesquisa elaborada por Grant e Booth (2009), esta se enquadra como "*scoping review*", categoria que tem como objetivo mapear conceitos e estruturas utillizadas em pesquisas dentro de determinada área do conhecimento.

A área de engenharia de ontologias e a criação de aplicações web semânticas tem expandindo muito nos últimos anos. No Figura 2 vemos o gráfico para os resultados para a pesquisa pelos termos "*open linked data platform*"no Google Scholar tendo como filtro secundário o ano de sua publicação. O crescimento de mais de 37% entre 2011 (218 mil resultados) e 2018 (299 mil) é ilustrativo do acentuado crescimento da área. Desconsidero aqui os anos entre 2019 e 2021 pela situação atípica em que a pandemia de Covid-19 nos colocou nesses anos. Outro fator que pode explicar a queda é a troca do termo "*open linked data platform*"por "*knowledge graphs*", expressão cujo uso vem se ampliando ao se designar a representação de uma rede de entidades do mundo real.

Figura 8 – Gráfico dos resultados da busca "*open linked data platform*"



Fonte: Autoria própria

Dentro do recente cenário pandêmico, aplicaçõs semânticas nos ajudam a construir

análises mais ricas e contextualizadas por fatores demográficos, temporais e geográficos. Na edição de 2020 da EU Datathon - evento organizado pela *Publications Office*[2] da União Européia que fomenta projetos que usam LOD de datasets existentes para endereçar questões sociais e econômicas - foi lançado um apanhado do uso dessas tecnologias no mapeamento da pandemia de COVID-19[3]. Dentre as iniciativas em destaque há o projeto CORD-19-on-FHIR, que tem como base os dados disponibilizados pelo projeto CORD-19[4] (COVID-19 *Open Research Dataset*) - o qual reúne publicações acadêmicas sobre COVID-19 para pesquisadores a nível global - e os transforma em LOD, acrescentando a eles anotações semânticas. Seu propósito é facilitar a conexão desses dados com grupos de dados biomédicos de outras fontes.

Diversas áreas vem se apropriando de aplicações web semânticas como meios de facilitar acesso à informação e gerar conhecimento. Nessa sessão citaremos três trabalhos publicados entre 2020 e 2021 que usaram de LOD - do inglês *Linked Open Data* - para criar plataformas que conectam informações antes dispersas, possibilitando a criação de novos conhecimentos e o tornando esse conteúdo acessível ao público em geral.

O projeto DIISH (do inglês *diet-improvement ingredient substitutability heuristic*) é um exemplo bastante ilustrativo do potencial prático do uso de dados conectados . Tal aplicação se faz valiosa para pessoas com restrições alimentares sérias - como alergias severas e diabetes - e para aquelas cujo objetivo é se direcionar a uma dieta mais sustentável e saudável (SHIRAI et al., 2021). Utilizando *Knowledge Graphs* (KG) construiu-se uma heurística de substituição de ingredientes que faz uso de informações semânticas conectadas disponibilizadas na Web, unindo três bases de dados públicas distintas e gerando novas informações nesse processo.

O trabalho utilizou informações nutricionais da USDA (*United States Department of Agriculture*), do KG semântico *FoodKG*, que conecta ingredientes de receitas e suas respectivas informações nutricionais, e da ontologia *FoodOn*[5], que categoriza os ingredientes e suas origens (SHIRAI et al., 2021). O projeto DISSH conecta essas três fontes de LOD, formando triplas que permitem encontrar receitas com um determinado ingrediente, a tabela nutricional e a proveniência do mesmo.

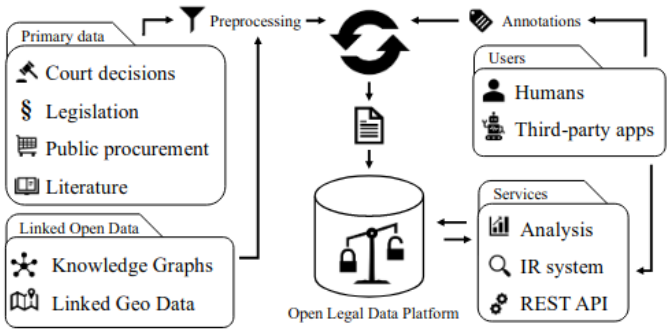
Um exemplo que tange o conceito de Dados Abertos Governamentais (DAG) é descrito no trabalho de Ostendorff, Blume e Ostendorff (2020). A plataforma *Open Legal Data Platform*[6] agrega dados relacionados a leis e processoas legais na Alemanha, para fomentar a transparência e a acessibilidade a dados públicos. A estrutura da aplicação desenvolvida é ilustrada na Figura 3. Ela fornece uma estrutura tecnológica básica que engenheiros jurídicos podem utilizar para desenvolver novas aplicações e adaptá-las conforme

[2] <https://op.europa.eu/en/home>
[3] <https://op.europa.eu/en/web/eudatathon/2020-edition>
[4] <https://www.semanticscholar.org/cord19>
[5] <https://foodon.org/>
[6] <https://openlegaldata.io/>

o ecossistema legal de cada país.

A aplicação *Open Legal Data Platform* se alimenta de fontes de dados primárias tais como legislações, autos legais e dados governamentais, e tem como fonte secundária dados abertos e conectados - como LOD de informações geográficas. No reprocessamento, dados são extraídos dos textos de fontes primárias, possibilitando pesquisas de texto estruturada e a conexão com os dados secundários na geração de informações adicionais. A platforma, por exemplo, pode extrair a localização de expedição de um determinado documento legal e o conectar aos dados da base LinkedGeoData[7]. Dentre os serviços que provê aos usuários estão a busca estruturada, o destaque automático de trechos de possível interesse nos documentos retornados e a visualização de relações entre os termos pesquisados e outros vocábulos, com ênfase nas relações mais frequentes.

Figura 9 – Estrutura da aplicação *Open Legal Data Platform*



Fonte: Ostendorff, Blume e Ostendorff (2020)

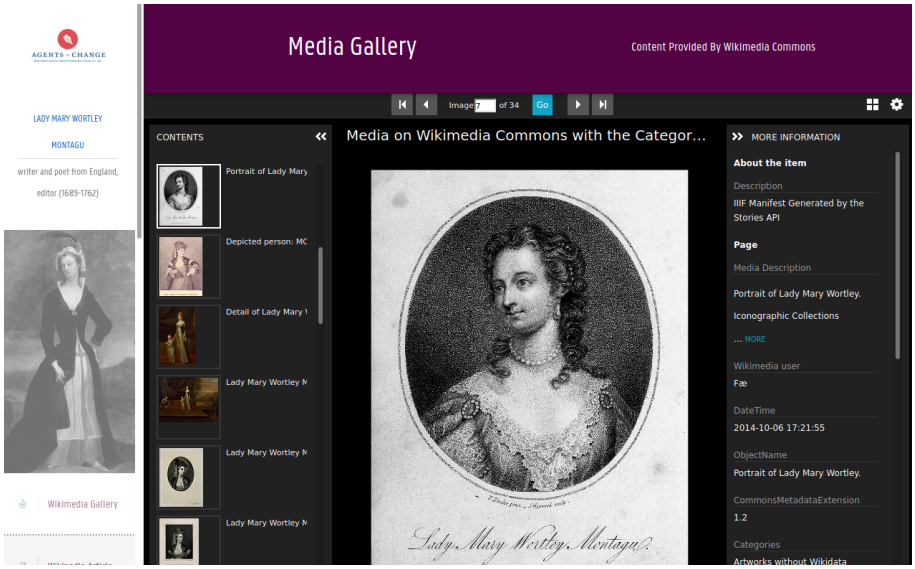
Um projeto que se aproxima mais do presente trabalho e relaciona Web Semântica e gênero é o *WeChangEd*, acrônimo para *Agents of Change: Women Editors and Socio-Cultural Transformation in Europe, 1710-1920*[8], desenvolvido entre 2015 e 2021 na *Ghent University*, Bélgica (THORNTON et al., 2021). O projeto se utilizou de um modelo de dados conectados para investigar o impacto de mulheres editoras entre 1710-1920, época em que mulheres eram podadas de diversos direitos fundamentais, incluindo-se o direito ao voto e o acesso à educação formal, na maior parte do mundo. Um dos principais resultados desse extenso trabalho é uma base de dados de mulheres editoras, seus periódicos, organizações e parcerias. Foram mapeadas aproximadamente 1700 pessoas, 1600 periódicos, 200 organizações e as numerosas conexões entre estas entidades (THORNTON et al., 2021). Os dados produzidos estão disponíveis publicamente na Wikidata[9], base de dados estruturados, abertos e conectados do movimento Wikimedia[10], cujo fruto mais conhecido

[7] <http://linkedgeodata.org/>
[8] <https://www.wechanged.ugent.be>
[9] <https://www.wikidata.org>
[10] <https://meta.wikimedia.org/wiki/Wikimedia_Foundation>

é a Wikipedia[11].

A plataforma resultante do trabalho de Thornton et al. (2021) conecta dados mapeados durante o projeto *WeChangEd* com dados previamente existentes na Wikidata na geração de *Stories* centrados na vida de mulheres editoras. Os *WeChangEd Stories* (Figura 4) são gerados automaticamente e destacam as publicações de autoria ou editadas por elas, lugares e pessoas relacionadas a elas, além de imagens e vídeos que se aproveitam da base de dados da Wikimedia Commons. A navegação se dá pelo menu lateral esquerdo - que direciona a usuária para as diversas sessões sobre a editora em questão - e inclui tanto links para outras entradas na plataforma quanto para outras websites.

Figura 10 – *WeChangEd Story* sobre Lady Mary Wortley Montagu



Fonte: *WeChangEd Stories* website

O formato interativo do site foi inspirado na plataforma *Science Stories* [12], projeto que conecta bases de dados de diversas instituições acadêmicas ao redor do mundo com bases de dados abertos com o intuito de dar visibilidade e tornar acessível a história das pessoas que construíram e constroem as ciências, especialmente as consideradas minorias socias (THORNTON; SEALS-NUTT, 2018). O *Science Stories* promove a presença da ciência em espaços sociais e demonstra o valor que agregar dados de acervos de diversas instituições pode gerar. A infraestrutura do projeto fomenta a acessibilidade e interatividade da ciência, além de servir de base para o desenvolvimento de projetos correlatos.

Imputar novos significados a dados antes dispersos através do potencial que sua conexão gera só é possível pela existência de uma rede mundial de informações. É no contextualizar de informações e em seu enriquecimento que reside o gigantesco potencial da web de verdadeiramente gerar conhecimento. Projetos como *DIISH*, *Open Legal Data*

[11] <https://en.wikipedia.org>
[12] <https://www.sciencestories.io>

Platform, *WeChangEd* e *Science Stories* se utilizam desse potencial para democratizar o acesso ao conhecimento, posicionando-se como base para projetos futuros, promovendo a elaboração de uma rede ainda mais rica e expressiva de dados.

O projeto ELLAS se alinha com os trabalhos aqui mencionados por ter como objetivo facilitar o acesso à informações sobre mulheres na área de STEM, fornecendo melhores ferramentas àqueles que tomam decisões. Para além disso se assemelham ao se construírem a partir de trabalhos e bases de dados já existentes, enaltecendo o propósito de conexão da web semântica.

Tabela 1 – Tabela comparativa de metodologias para elaboração de ontologias

	101	**AMOD**	**NeOn**	**SABiO 2.0**
Colaborativi dade	Baixa, depende da ferramenta	Baixa, depende da ferramenta	Alta, integrada ao desenvolvimento	Alta, integrada ao desenvolvimento. Sugere o uso de Wikis
Conectividade	Alta, cria estruturas de dados em tuplas (RDF)	Alta, cria estruturas de dados em tuplas (RDF)	Alta, cria estruturas de dados em tuplas (RDF)	Alta, cria estruturas de dados em tuplas (RDF)
Reusabilidade	Alta, integrada ao desenvolvimento	Alta, integrada ao desenvolvimento	Alta, integrada ao desenvolvimento	Alta, integrada ao desenvolvimento
Etapas	Abordagem "top-down": 1. Determinar o domínio e o escopo da ontologia 2. Considerar reutilizar ontologias já existentes 3. Identificar termos importantes na ontologia 4. Definir as classes e suas hierarquias 5. Definir as propriedades das classes 6. Definir os aspectos das propriedades 7. Criar instâncias	Etapas cíclicas: 1. Pré-jogo: definição de escopo, objetivo, requisitos, técnicas e ferramentes, seleção de fontes 2. Desenvolvimento: planejamento da sprint, aquisição de conhecimento, integração e conceitualização, formalização e revisão da sprint 3. Pós-jogo: avaliação, manutenção Atividades de suporte: documentação e gerenciamento de configuração	Variável conforme o cenário estabelecido, estando sempre presentes: 1. Aquisição de conhecimento 2. Documentação 3. Gerenciamento de configuração 4. Avaliação	Delimita 6 papéis bem definidos, tendo como etapas: 1. Identificação de propósito e requisitos 2. Captura e formalização da ontologia (1 & 2 são iterativos) 3. Design 4. Implementação 5. Testes Atividades de suporte: aquisição de conhecimento, documentação, gerenciamento de configuração, avaliação, reuso
Documentação	Elaboração não definida	Recorrente e inclusa como atividade de suporte	Elaboração mencionada mas não definida	Recorrente e inclusa como atividade de suporte
Ferramentas	Protégé, Ontolingua e Chimaera	Protégé	Protégé	OntoUML, OLED, Protégé

Fonte: Autoria própria

3 METODOLOGIA

Este capítulo apresenta a metodologia proposta para que a presente pesquisa alcance os resultados e os objetivos propostos nos capítulos anteriores. Serão apresentadas as Etapas de Desenvolvimento e as Etapas da Metodologia ELLAS.

Por se situar num contexto de estudo de características de um grupo (estudantes de computação no Brasil) e ter aplicação específica (entender as dinâmicas e fatores ao redor de gênero na computação no Brasil), consideradas as classificações possíveis para pesquisas explicitadas por Gil (), o presente trabalho é definido como uma pesquisa aplicada e descritiva. Podemos-mos também definí-la como uma pesquisa mista entre qualitativa e quantitativa, de métodos mistos entre pesquisa bibliográfica e pesquisa-ação - por ser situacional e ter como finalidade conduzir à ação social.

inalidades pos tipo de pesquisa e abordagem

3.1 Etapas de Desenvolvimento

1. Definição do escopo: análise e delimitação de escopo do trabalho dentro dos objetivos do projeto ELLAS.
2. Referencial teórico: revisão de literatura na área de engenharia de ontologias, apresentação de conceitos que embasam a construção do trabalho e levantamento de trabalhos correlatos, com ênfase em pesquisas sobre o uso de dados abertos conectados.
3. Desenvolvimento de uma metodologia para elaboração de ontologias: especificação, com base em metodologias já existentes, de uma metodologia reutilizável para o projeto ELLAS.
4. Desenvolvimento da ontologia: instanciação da metodologia para elaboração de ontologias através de processos de aquisição de conhecimento, especificação, conceitualização e finalmente a formalização da ontologia de referência.
5. Conclusão do trabalho: avaliação do modelo ontológico desenvolvido e de seu potencial de escalabilidade e colaboratividade, com foco em sua implementação nas fases seguintes do ELLAS.

Dado o levantamento realizado na seção 2.4 e a natureza do projeto ELLAS, como metodologia para elaboração de ontologias optou-se por uma customização com elementos das metodologias NeOn e SABio 2.0, aproveitando-se da perspectiva de reutilização de recursos não-ontológicos da primeira e da ênfase em colaborativitidade e documentação da última. A metodologia será referenciada como Metodologia ELLAS e tem com objetivo ser utilizada por diversos grupos de trabalho dentro do projeto, sendo o presente trabalho uma instanciação da abstração apresentada na Metodologia ELLAS aplicada ao contexto brasileiro.

3.2 Etapas da Metodologia ELLAS

Na seção 2.4 observamos que todas as metodologias estudadas (101, AMOD, NeOn E SABiO 2.0) convergem ao estabelecer que nas fases iniciais da elaboração de uma ontologia estão a determinação de seu escopo e a aquisição de conhecimento - atividade que inclui a análise e seleção de recursos para reutilização. Outra atividade importante dentro das fases iniciais - dentro da definição de escopo ou de requisitos - é a descrição das Questões de Competência (QCs).

Elas são o conjunto de perguntas que uma ontologia, ou um conjunto delas, deve ser capaz de responder usando o conhecimento representado por seus axiomas (GRUNINGER; FOX, 1995). Assim, uma lista de QCs está para as ontologias como a especificação de requisitos funcionais está para o desenvolvimento de software. Seu objetivo é verificar se a ontologia foi criada corretamente, ou seja, se contém os axiomas necessários para responder as questões propostas.

Conforme estabelecido na sessão 2.4, é também nos passos inicias que a metodologia SABiO estabelece os papéis necessários na elaboração de uma ontologia. Determinamos as funções a serem utilizadas na metodologia ELLAS a partir deles, a constar: (1) especialista de domínio, que fornece o conhecimento a ser modelado na ontologia de domínio; (2) usuária/o da ontologia; (3) engenheira/designer de ontologias, responsável pelas ontologia de referência e design da ontologia operacional e (4) programadora/testadora de ontologias, a implementar a ontologia operacional e testá-la. Assim definimos o a primeira etapa da Metodologia ELLAS:

- 1. Definição do escopo e divisão de papéis: determinação do objetivo e do domínio da ontologia, especificação dos requisitos (QCs), assim como dos papéis a serem assumidos pela equipe.

Como o cenário do projeto ELLAS inclui a reutilização de recursos, sua seleção é um sucessor natural do primeiro passo, se alinhando ao cenário 2 da NeOn e a passos iniciais das metodologias 101 e AMOD. Definimos então o segundo passo:

- 2. Seleção de recursos: pesquisa, avaliação e seleção de recursos relacionados ao domínio - ontológicos e não-ontológicos - para reutilização na ontologia.

Em seguida, temos a atividade de reestruturação dos recursos, ontológicos e não-ontológicos, para reuso. Tal movimento visa transformar os componentes mais relevantes dos recursos selecionados em um modelo conceitual e se alinha a passos descritos nas 4 metodologias analisadas. Assim, estabelecemos a próxima fase:

- 3. Reestruturação de recursos: limpeza, reorganização e estruturação dos recursos selecionados para reuso no modelo conceitual da ontologia.

No fluxo especificado pela AMOD, temos integração, conceitualização e formalização após as fases inicias. Conforme a Figura 7 explicita, no cenário 2 da NeOn temos a

conceitualização da ontologia, seguida de sua formalização. Já na linearidade da SABiO 2.0, temos captura e formalização. Todas convergem ao estabelecer a documentação como parte importante da metodologia, logo estabelecemos como etapa seguinte:

- 4. Especificação e conceitualização da ontologia: elaboração da ontologia de referência - com a definição de suas classes, propriedades e hierarquias - e elaboração de um Dicionário da Ontologia que as traduzam.

Dentre as metodologias pesquisadas, todas convergem quanto a instanciação/implementação da ontologia, sendo que AMOD, NeOn e SABio convergem no quesito avaliação. À vista disso, temos como fases finais:

- 5. Instanciação da ontologia: transformar a ontologia de referência em operacional através de software de gerenciamento de conhecimento.
- 6. Avaliação da ontologia: aferir o potencial da ontologia criada de fornecer respostas às Questões de Compentência sobre mulheres nas áreas de STEM nos países participantes do ELLAS.

O software Protégé[1] - um editor de ontologias e um sistema de gerenciamento de conhecimento criado em 1987, mantido pela universidade de Stanford - foi a ferramenta eleita para a elaboração da ontologia. O Protégé é uma ferramenta gratuita, de código aberto e amplamente utilizada na área de engenharia de ontologias, contando com vasta documentação e diversos plugins que facilitam seu uso. Para a instanciação da ontologia, utilizamos o software Ontotex GraphDB[2] - um banco de dados gráfico e uma ferramenta de descoberta de conhecimento compatível com RDF e SPARQL que se alinha aos padrões W3C. Nela também transformamos as bases de dados relacionais (CSV) em triplas RDF, função baseada na aplicação OpenRefine[3]

[1] <https://protege.stanford.edu/>
[2] <https://graphdb.ontotext.com/documentation/10.1/about-graphdb.html>.
[3] <https://openrefine.org/>

4 DESENVOLVIMENTO

O presente capítulo descreve, no contexto brasileiro, uma instanciação da Metodologia ELLAS, proposta na seção 3.2 e cujo intuito é ser utilizada em instanciações diversas dentro do Projeto ELLAS.

4.1 Definição do Escopo

Como recorte do escopo do projeto ELLAS, o presente trabalho tem foco em gênero e computação - o T de *Technology* da sigla STEM. O ELLAS tem o objetivo de conglomerar e organizar dados e informações que tangem a presença de mulheres em posições de liderança nas áreas de STEM na América Latina - inicialmente Brasil, Peru e Bolívia. No presente trabalho temos um recorte desse escopo para o Brasil, e os recursos a serem analisados são especificamente no espectro educacional da computação.

O ponto de partida escolhido para a formulação das questões de competência - e também como recursos não ontológicos a serem reutilizados - foram os relatórios publicados pela Sociedade Brasileira de Computação (SBC). Temos como objetivo construir uma ontologia que permita responder questões como:

- Quantas mulheres ingressaram em cursos de computação no Brasil em determinado ano? E em determinada região/estado/cidade do país?
- Quantas mulheres concluíram bacharelados em cursos de computação em universidades públicas no Brasil em determinado ano?
- Entre concluintes dos cursos de Sistemas de Informação em capitais do país em 2021, qual a porcentagem de mulheres?
- Quantas mulheres estavam matriculadas em Institutos Federais de Educação, Ciência e Tecnologia na área da computação no Nordeste em determinado ano?

Tais perguntas são importantes pois suas respostas nos permitem construir um panorama embasado em dados sobre as mulheres nos cursos de computação no Brasil - panorama este que até o momento se constrói através de estudos de caso e percepções menos objetivas. Computar dados de todos os cursos do país da área de Computação e TCI também nos permite elaborar comparativos diversos, por exemplo visualizando as mudanças ao longo dos anos, comparando os números entre estados com índices de desenvolvimento humano distintos ou entre cidades com mais ou menos creches, com melhores ou piores índices de violência contra a mulher.

Responder a essas perguntas com dados dá peso aos argumentos usados, coloca a realidade apontada como mais concreta para todos aqueles que não a vivem - fator necessário na construção de políticas públicas que englobem a todas e todos na sociedade. Além disso são informações objetivas e mensuráveis para realizar comporativos com a

realidade de outros países e fomentar análises interseccionais. A facilidade na extração de significados e a possibilidade de conexão com outras bases de dados que uma ontologia oferece destaca o presente trabalho dos dados que o Censo da Educação Superior do Inep já nos oferece.

Importante também considerar que há questões cujas respostas - embora relevantes dentro do contexto do projeto ELLAS e para além deste - não conseguimos obter por limitações da fonte de dados. Entram nessa categoria a interseção de gênero com raça, idade, deficiências e programas educacionais - como o Prouni e o FIES[1]. Também pelas limitações dos dados não é possível responder perguntas sobre a presença de mulheres no quadro docente de cursos superiores nas áreas da computação no Brasil.

4.2 Seleção de Recursos

Há pouca informação sobre questões de gênero no Brasil e no aspecto ocupacional as fontes de informação são ainda mais escassas. Mesmo que não fosse esse o contexto, seria apenas natural procurar a Sociedade Brasileira de Computação (SBC), a maior sociedade científica da área do país, como referência.

Os relatórios da SBC são publicados em formato PDF e organizam informações sobre os cursos relacionados à computação no Brasil em gráficos, inviabilizando uma análise exata dos números e dificultando seu reuso. Ao contactar o orgão, fomos informados que os relatórios são baseados nos microdados do Censo da Educação Superior. O Censo é o instrumento de pesquisa mais completo do Brasil sobre as instituições de educação superior (IES) que ofertam cursos de graduação e sequencias de formação específica, além de seus alunos e docentes. O objetivo da coleta é oferecer informações estatísticas confiáveis, que permitam conhecer e acompanhar o sistema brasileiro de educação superior[2]. É realizado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) desde 1995, e desde 2009 publica dados em formato CSV.

A divulgação dos dados em CSV faz do Censo uma fonte segura e eficaz de obtenção de dados, acessível a pesquisadores, estudantes, gestores e sociedade em geral e através deles é possível obter um amplo panorama da educação brasileira. A alta conectividade e a facilidade de extrair informações específicas desses dados faz deles fonte excepcionalmente adequada aos objetivos deste trabalho, além de possibilitar a elaboração de Questões de Competência.

[1] Prouni e FIES são programas educacionais afirmativos que visam facilitar o acesso à educação superior no Brasil. O Programa Universidade Para Todos (Prouni) oferta bolsas de estudo, integrais e parciais em cursos de graduação e sequenciais de formação específica, em instituições de educação superior privadas, já o Fundo de Financiamento Estudantil (FIES) é um programa que financia a graduação de estudantes matriculados em cursos superiores não gratuitos. Ambos são programas do Ministério da Educação.
<https://acessounico.mec.gov.br/prouni/duvidas#o-prouni>
<https://www.caixa.gov.br/programas-sociais/fies/Paginas/default.aspx>

[2] <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior>

Trabalhamos com o Censo de Educação Superior de 2021, publicado em novembro de 2022. Ele contém duas grandes tabelas com dados em formato CSV, o dicionário de dados e os questionários utilizados. Há a tabela Microdados Cadastro Cursos e a tabela Microdados Cadastro IES, sendo que a última não contém dados que nos permitam fazer um recorte entre gênero e computação a nível institucional. Desta utilizamos as colunas com informações relevantes a respeito da Instituição de Ensino Superior, como Rede de Ensino (Pública ou Privada) e Categoria Administrativa (Federal, Estadual, Privada sem fins lucrativos...).

4.2.1 Plataformas Relacionadas

No mapeamento de entidades não-ontológicas relacionadas ao projeto pesquisamos plataformas que se utilizam de dados abertos no endereçamento de questões sociais atreladas à tecnologia. Tal pesquisa resultou em três resultados principais:

1. GoSPIN - *Global Observatory of Science, Technology and Innovation Policy Instruments*[3]: plataforma da UNESCO de acesso aberto que oferece ferramentas analíticas para mapeamento e análise de políticas relacionadas à *science, technology and innovation* (STI) a nível global
2. EC-OECD STIP Compass[4]: iniciativa da *European Commission* (EC) e da *Organisation for Economic Co-operation and Development*[5] (OECD) que monitora e analisa dados qualitativos e quantitativos sobre políticas de STI a nível global, com dados abertos e acessíveis senguindo os princípios FAIR (*Findable, Accessible, Interoperable, and Re-usable*)
3. SAGA - *STEM and Gender Advancement*[6]: base de dados, desenvolvida entre 2015 e 2018 pelo governo da Suécia e pela UNESCO, de políticas que endereçam a desigualdade de gênero nas áreas de STI em mais de 50 países.

Tais plataformas se mostram relevantes ao projeto ELLAS e deverão ser consideradas para o enriquecimento da ontologia. Contudo decidimos focar nos microdados do Inep - fonte muito rica sobre a educação superior no Brasil. Devido à sua dimensão - a tabela Microdados Cadastro Cursos contém 200 colunas e quase 445 mil linhas e a de dados das IES mais de 2,5 mil instituições e 58 colunas - este foi o único recurso selecionado para o presente trabalho.

4.3 Reestruturação de recursos

Na limpeza dos recursos o primeiro filtro aplicado aos dados do Inep foi delimitar os cursos apenas àqueles dentro da Área Geral de Computação e Tecnologias da Informação e

[3] <https://gospin.unesco.org/frontend/StatPlanet_Cloud/init.php>
[4] <https://stip.oecd.org/stip/>
[5] <https://www.oecd.org/>
[6] <https://en.unesco.org/saga>

Comunicação (TIC). As classificações de Área Geral e Área Detalhada utilizadas pelo Inep seguem a Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais (Cine Brasil), criada pelo orgão tendo como referência a metodologia da *International Standard Classification of Education – Fields of Education and Training* (ISCED-F 2013), com adequações à realidade educacional brasileira. A ISCED é produzida pelo Instituto de Estatísticas da Unesco com a finalidade de reunir, compilar e analisar estatísticas educacionais comparáveis com as de outros países, regiões e estados, e possibilita ordenar cursos por níveis de ensino e áreas de formação[7].

Integramos as colunas relevantes em apenas um arquivo CSV e para facilitar a manipulação de dados. No processo de tratamento de dados utilizamos pesquisas verticais[8] para conectar as informações dos cursos com as das IES, contidas em tabelas distintas. Escolhemos traduzir os nomes das colunas para inglês pelo consenso sobre sua utilização na área e pela natureza internacional do projeto ELLAS - uso que se estende na conceitualização da ontologia. Uma coluna, is_free, inexistente no Censo de 2020, foi adicionada e informa se um determinado curso é gratuito ou não.

Como resultado desse processo criamos a tabela Censo da Educação Superior - Comp e TIC, formada por 30 colunas e quase 40 mil linhas - referentes apenas ao ano de 2021. Nesta planilha encontramos algumas entradas classificadas como ensino remoto e cujos campos de localização estavam vazios. Por serem dissonantes do restante dos dados e somarem apenas 30 linhas, decidimos removê-las. Deletamos também entradas de cursos que contavam com 0 alunos matriculados.

4.4 Especificação e Conceitualização da Ontologia

A primeira versão da Ontologia ELLAS-CompBRA foi baseada nos dados do Inep de 2020 e nela estabelecemos o uso dos dois arquivos CSV - relativos a Curso e IES - e criamos conexão entre eles. A princípio consideramos o uso de um mesmo predicado para estabelecer as relações entre as classes e literais, <pertence_a> e <contém>, sendo inversos um do outro - por exemplo, teríamos as triplas (<curso>, <pertence a>, <IES>) e (<IES>, <contém>, <curso>). Havíamos estabelecido a classe <localização> como concatenação de <região>, <estado> e <município>, tendo estas como subclasses e pensado nos literais com os números de matrículas totais, ingressantes e concluintes em dois níveis - o total e como subclasses deles estariam a divisão entre sexos feminino e masculino.

Ao longo do processo de implementação, todos esses pontos foram revistos. Os predicados foram renomeados de forma mais descritiva, de modo a facilitar a identificação

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

[8] Função que procura um valor específico em uma tabela e retorna, em uma nova tabela, o valor correspondente que existe na mesma linha, mas em uma coluna diferente. Também conhecido como PROCV ou VLOOKUP.

de sujeito e objeto. A hierarquia tanto das subclasses de <localização> quanto dos literais das estatísticas foram retirados da ontologia - o primeiro pela classe <localização> ter sido retirada da ontologia na tentativa de minimizar possíveis fontes de erro no mapeamento e o segundo por causar dados em duplicidade.

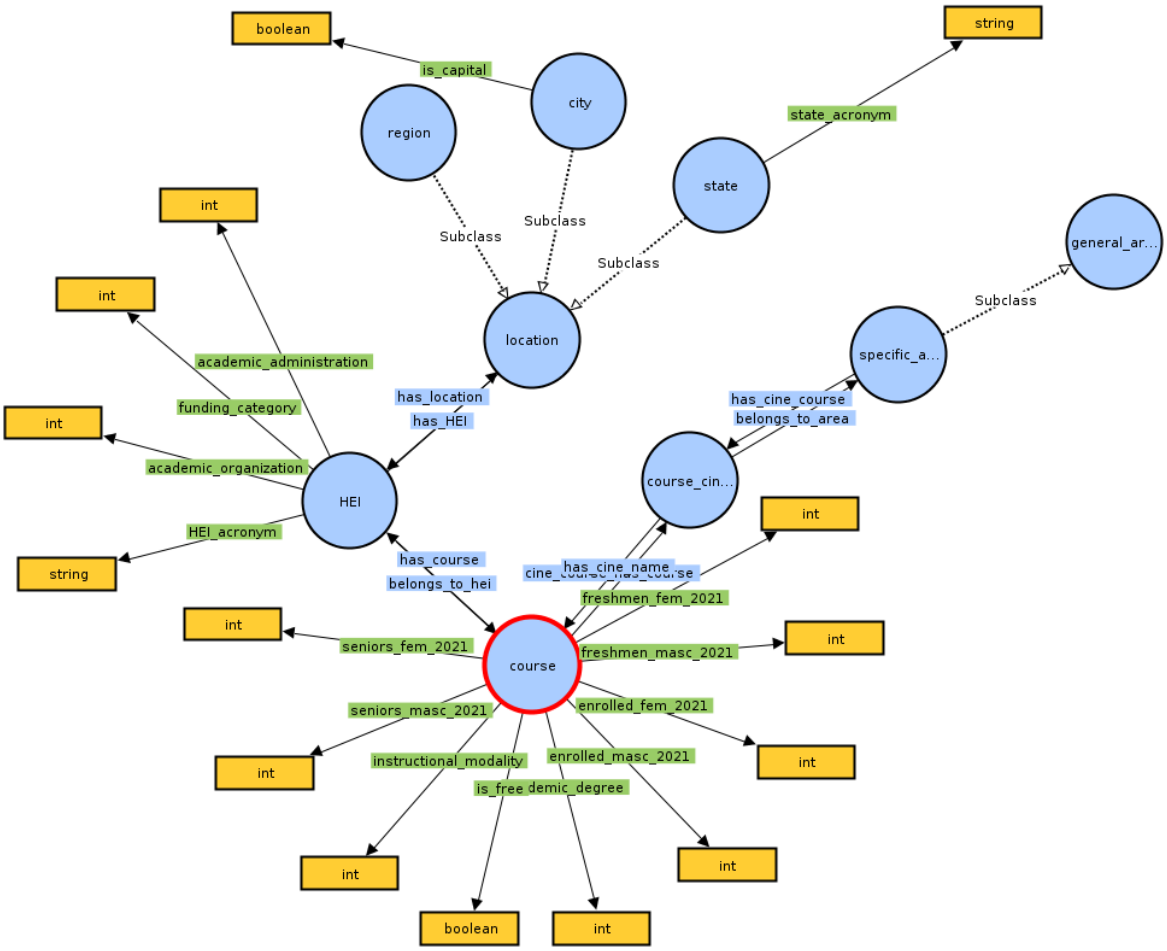
A ontologia de referência foi modelada utilizando-se o software Protégé. Na Figura 11, é possível visualizar a Ontologia ELLAS-CompBRA em sua primeira versão no Protégé através do VOWL - que implementa o *Visual Notation for OWL Ontologies*[9]. Nela temos 6 classes, 3 subclasses, 8 propriedades de objetos - que conectam classes - e 14 propriedades de dados - que conectam classes a literais. Os processos de conceitualização e de instanciação da ontologia aconteceram de forma iterativa - falhas de conceitualização foram identificadas na instanciação e o modelo foi recorrentemente corrigido até que as consultas SPARQL retornassem resultados consistentes.

A falha central na conceitualização inicial foi a ausência de um identificador único para cada entrada - o que causava estatísticas pertencentes a diversas instâncias a estarem ligadas ao mesmo nome de curso. Após a correção no mapeamento com a criação da classe *<id_course>* como elemento identificador - e as consequentes alterações em propriedades de dados e de objetos - refinar a ontologia se resumiu a questões de instanciação.

Na Tabela 2 temos a lista das propriedades de objetos revisada, com as 16 propriedades que conectam classes na ontologia. Já na Tabela 3 vemos a relação final das 15 propriedades de dados concebidas, conectando classes a literais. Foi também elaborado um Dicionário de Dados, que reúne os termos utilizados na descrição dos objetos modelados para a ontologia. Seu objetivo é explicitar qual é a estrutura lógica da ontologia e ele pode ser visto no Apêndice B. Nele também se encontram a descrição completa completa das classes e das propriedades de dados.

[9] <http://vowl.visualdataweb.org/protegevowl.html>

Figura 11 – Primeira versão da Ontologia ELLAS-CompBRA no Protégé



Fonte: Autoria própria

Tabela 2 – Propriedade de objetos da Ontologia ELLAS-CompBRA

Propriedade de objetos	**Domínio**	**Imagem**
course_belongs_id	course	id_course
belongs_to_area	course_cine	specific_area
cine_has_course	course_cine	course
belongs_to_gen_area	specific_area	general_area
contains_spec_area	general_area	specific_area
has_cine_course	specific_area	course_cine
has_cine_name	course	course_cine
hei_belongs_id	hei	id_course
id_has_course	id_course	course
id_has_hei	id_course	hei
id_city	id_course	city
id_state	id_course	state
id_region	id_course	region
c_has_id	city	id_course
s_has_id	state	id_course
r_has_id	region	id_course

Tabela 3 – Propriedades de dados da Ontologia ELLAS-CompBRA

Propriedade de dados	**Domínio**	**Imagem**
is_capital	city	Inteiro
academic_administration	hei	Inteiro
academic_organization	hei	Inteiro
funding_category	hei	Inteiro
hei_acronym	hei	String
academic_degree	id_course	Inteiro
instructional_modality	id_course	Inteiro
is_free	id_course	Inteiro
enrolled_fem_2021	id_course	Inteiro
enrolled_masc_2021	id_course	Inteiro
freshmen_fem_2021	id_course	Inteiro
freshmen_masc_2021	id_course	Inteiro
seniors_fem_2021	id_course	Inteiro
seniors_masc_2021	id_course	Inteiro
state_acronym	state	String

Fonte: Autoria própria

4.5 Instanciação da Ontologia

A instanciação se deu no software Ontotex GraphDB, onde mapeamos as relações entre as colunas, tanto as que se instanciam classes quanto as que instanciam propriedades de objetos (Figura 12).

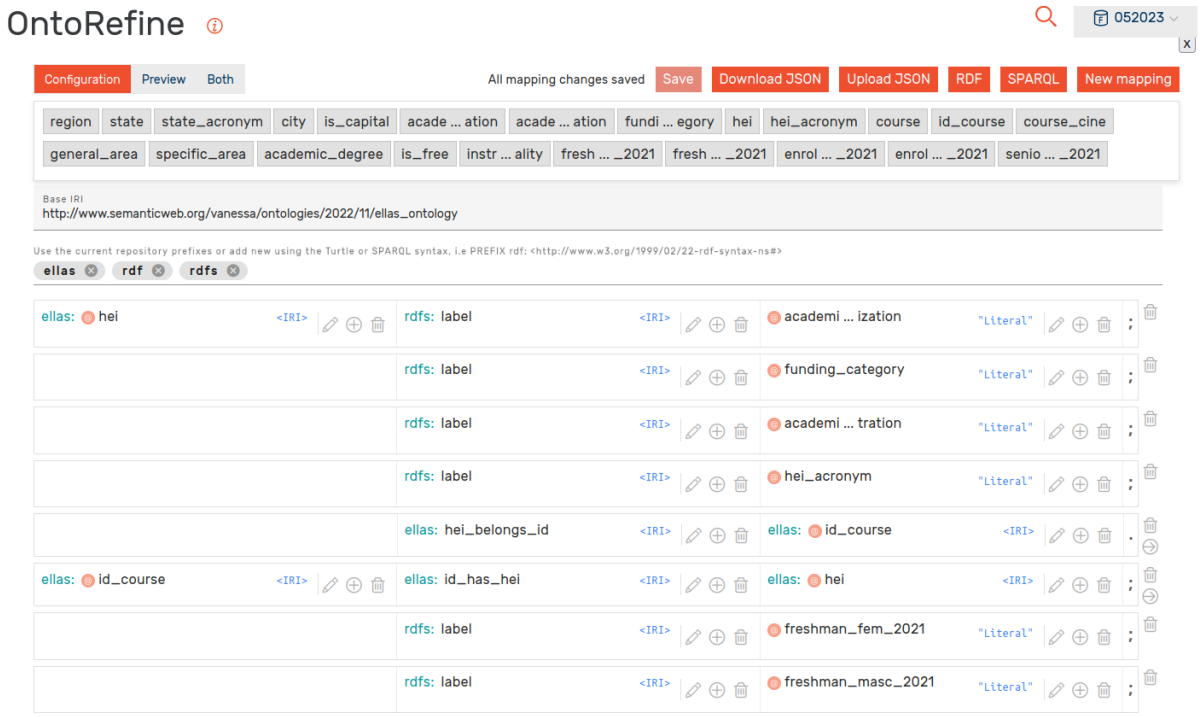
A primeira tentativa se baseou na modelagem da Figura 11, no entanto a criação de triplas não aconteceu de forma adequada. Para mitigar o problema criamos um identificador único, <id_course>, concatenação das colunas course, enrolled_total[10], hei_acronym e city, ao qual atrelamos as estatísticas daquele curso específico. Também deletamos a classe *<location>*, atrelando *<region>*, *<state>* e *<city>* diretamente a classe *<id_course>*. Foram criadas também relações entre as classes *<region>*, *<state>* e *<city>*, explicitando que cidades são contidas em estados, estados pertencem a regiões específicas e contém cidades e regiões contêm estados. A versão final do modelo de referência pode ser visto na Figura 13.

Outro processo importante dentro da instanciação da ontologia foi a normalização dos dados, bastante simples pelo fato da base do Inep já ser padronizada. Foram retirados caracteres especiais do arquivo CSV e todo seu conteúdo foi deixado em caixa alta para facilitar as consultas.

Após as alterações, foi gerado um arquivo com mais de 614 mil triplas. O mapeamento do arquivo CSV do Inep para as triplas em RDF foi feito através da ferramenta

[10]Coluna existente nos dados do Censo de Educação Superior do Inep e utilizada em <id_course> mas que não faz parte da Ontologia ELLAS-CompBRA

Figura 12 – Captura de tela do mapeamento no OntoRefine



Fonte: Autoria própria

OntoRefine[11] dentro do GraphDB[12].

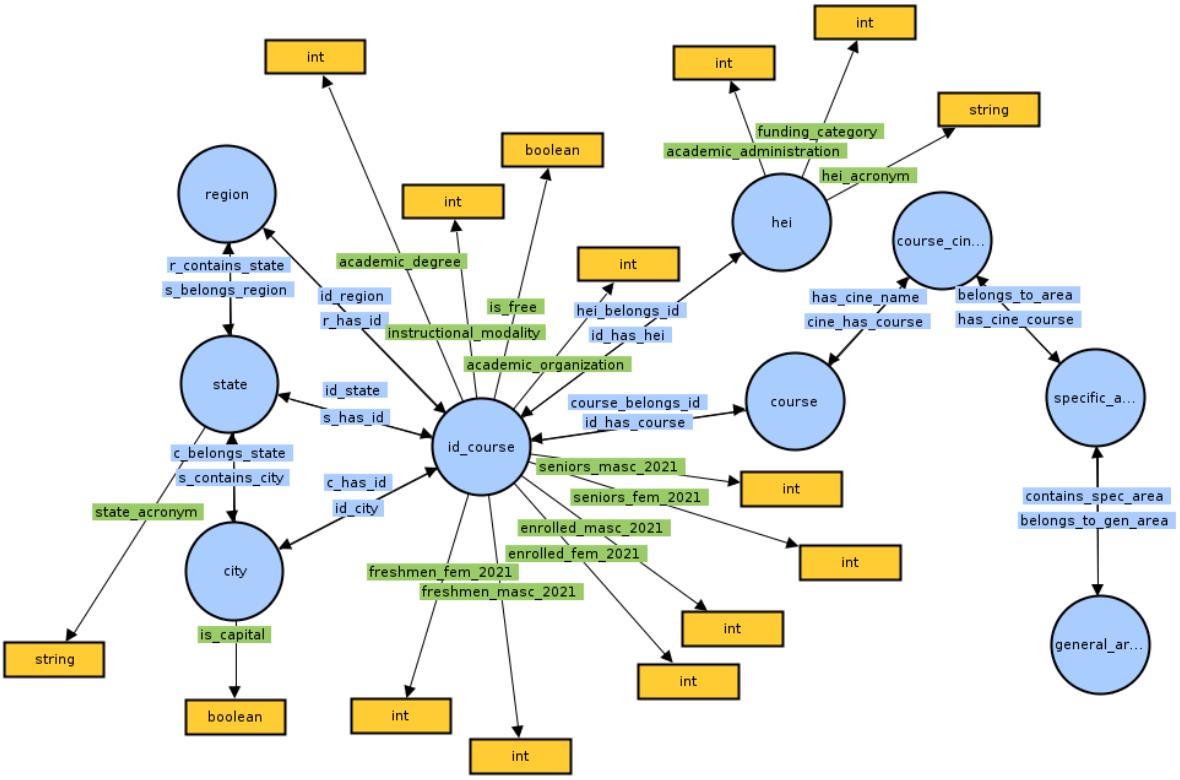
4.6 Avaliação da Ontologia

A avaliação da ontologia ELLAS-CompBRA se deu através da sua capacidade de responder as Questões de Competência estabelecidas na sessão 4.1. Consultas em SPARQL foram elaboradas para validar cada uma delas, sendo a primeira pergunta (QC1): "Quantas mulheres ingressaram em cursos de computação no Brasil em determinado ano?". Considerada a fonte de dados, utlizamos como ano 2021 para as consultas, assim podemos ver a captura de tela do GraphDB para o algoritmo elaborado para esta pergunta na Figura 14. A resposta, 50704, foi validada através do CSV e abaixo temos a lista com todas as Questões de Competência delineadas e suas respectivas respostas. Para as demais QCs também foram obtidos resultados que se confirmaram corretos, provando assim que esta instanciação da ontologia é funcional e seus resultados assertivos. As consultas completas em SPARQL utilizados para as consultas se encontram no Apêndice C.

[11] Ferramenta gratuita e open source que transforma banco de dados relacionais em triplas de formato RDF. <https://www.ontotext.com/products/ontotext-refine/>.

[12] OntoRefine é uma ferramenta independente do GraphDB, mas há neste uma integração com o OntoRefine, facilitando o processo de mapeamento e consulta às bases de dados da ontologia.

Figura 13 – Versão final da Ontologia ELLAS-CompBRA



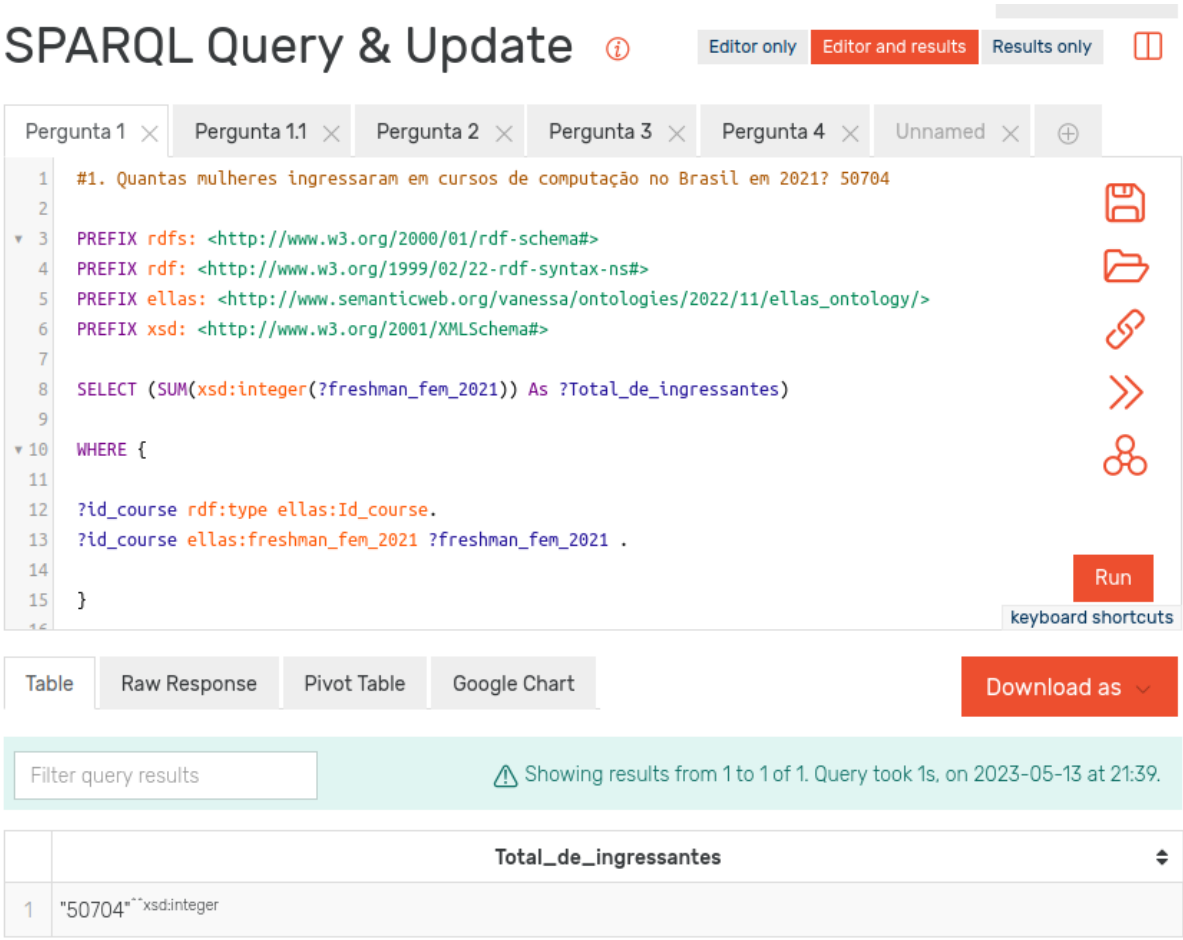
Fonte: Autoria própria

- (QC1) Quantas mulheres ingressaram em cursos de computação no Brasil em 2021?
  Resposta: 50704
- (QC1.1) Quantas mulheres ingressaram em cursos de computação em Curitiba em 2021?
  Resposta: 1391
- (QC2) Quantas mulheres concluíram bacharelados em cursos de computação em universidades públicas no Brasil em 2021?
  Resposta: 737
- (QC3) Entre concluintes dos cursos de Sistemas de Informação em capitais do país em 2021, qual a porcentagem de mulheres?
  Resposta: 14,198%
- (QC4) Quantas mulheres estavam matriculadas em Institutos Federais de Educação, Ciência e Tecnologia na área da computação no Nordeste em 2021?
  Resposta: 1662

Ao validar a capacidade desta instanciação da ontologia de responder às Questões

de Competência propostas, validamos também a efetividade da Metodologia ELLAS para a elaboração de ontologias que respondam satisfatoriamente a questões relacionadas à presença de mulheres nas áreas de STEM. Os requisitos estabelecidos em forma de Questões de Competência foram inteiramente atendidos, ratificando a eficácia da ontologia implementada no contexto brasileiro.

Figura 14 – Consulta SPARQL para a QC 1 no GraphDB



Fonte: Autoria própria

5 CONSIDERAÇÕES FINAIS

Este documento apresentou a pesquisa para o trabalho de conclusão de curso em Bacharelado em Sistemas de Informação intitulado "Elaboração de uma ontologia que endereça a presença de mulheres nos cursos de computação no Brasil". A pesquisa teve como pontos principais em sua trajetória a contextualização do problema e da área, a análise de trabalhos correlatos, a especificação da metodologia do trabalho - sendo criada uma metodologia própria para elaboração de ontologias, a Metodologia ELLAS - e a instanciação desta na contrução da Ontologia ELLAS-CompBRA. Nesta instanciação foram reestruturados dados do Inep à nível Brasil, contudo no contexto do ELLAS teremos variações - o que também nos levará a outras ontologias e nomenclaturas.

De início, o pouco conhecimento das ferramentas e processos relacionados à instanciação de ontologias e bancos de dados semânticos se colocou como um desafio para o desenvolvimento da pesquisa. Este desafio foi superado ao longo da evolução do trabalho e contou com a colaboração de outros pesquisadores do projeto. Por fim a presente instanciação da Metodologia ELLAS, realizada com dados do contexto da educação superior no Brasil, se mostrou bem sucedida.

Ao se considerar os objetivos geral e específicos do trabalho, sessão 1.1, vemos que foram atingidos satisfatoriamente. Foi construída e testada uma ontologia - uma estrutura da qual é possível extrair sentido - que possibilita delinear um panorama sobre a presença e permanência de mulheres na área da computação no Brasil à nível acadêmico. Outra importante contribuição é a Metodologia ELLAS, que já está sendo utilizada em outros ramos do projeto e cujas implementações hão de abranger outras fontes de informação, outros domínios, outros contextos e outros países.

Ainda assim existem diversos pontos de melhorias. O maior deles se localiza dentro da reestruturação de recursos, sessão 4.3, onde lidamos com os dados no Inep. Utilizamos de forma literal os dados de algumas colunas em que números inteiros foram utilizados para o preenchimento de campos - mas cujo significado é somente explicitado ao se consultar o seu dicionário de dados. Por exemplo, o campo <tp_grau_academico>, que especifica o grau acadêmico conferido a quem conclui o curso e que foi mapeado na ontologia como <academic_degree>, pode ser preenchido com números inteiros de 1 a 4. Ao se consultar o dicionário de dados vemos que 1 significa "Bacharelado", 2 significa "Licenciatura", 3 significa "Tecnólogo"e 4 significa "Bacharelado e Licenciatura- contudo o acesso ao dicionário é necessário para a compreensão dessas informações.

Para a extração mais direta da semântica da ontologia, os valores inteiros dos dados de entrada precisam ser convertidos em seus significados reais, eliminando a necessidade de uma consulta ao seu dicionário de dados. Nesse processo, algumas propriedades de dados se tornariam classes por serem entidades complexas, através das quais outras

ontologias poderiam se conectar. Por exemplo, <academic_degree> poderia se conectar a uma ontologia que mapeasse dados educacionais, explicitando relações como o título que determinado <academic_degree> proporciona ou quantos anos de formação tem um específico <academic_degree>.

As propriedades de objetos <academic_administration>, <instructional_modality>, <academic_organization> e <funding_category> também herdaram do CSV o preenchimento com números inteiros cuja compreensão depende do dicionário de dados. Aqui a habitual mentalidade de pensar nos dados como entidades de um modelo relacional se sobressaiu a percepção mais semântica sobre eles, e do seu entedimento como entidades abertas e conectadas - premissas centrais na elaboração de ontologias.

Utilizando a ontologia construída, podemos comparar dados relativos às alunas e alunos de cursos de computação no país por localização geográfica, por tipo de curso e instituição, por área da computação. A granularidade aqui estabelecida é - até onde se pesquisou - inédita, e maior do que a observada em trabalhos similares. Os relatórios anuais da SBC - que utilizam os mesmos dados do Censo do Inep - mostram um panorama geral sem a possibilidade ser mais específico ou de conectividade.

O potencial deste trabalho é ampliado ao ser conectado a ontologias e dados de outras atividades do ELLAS - como a mapeamento de iniciativas e políticas públicas e privadas que tangem a presença de mulheres em STEM nos três países do projeto, trabalho já em andamento em outras frentes do ELLAS. O resultado de todos os trabalhos do ELLAS culminarão numa plataforma de dados abertos conectados sobre gênero e STEM, senda esta uma das pesquisas que ajuda a entender esse contexto.

O trabalho futuro mais relevante a se mencionar é a correção das propriedades de dados que necessitam de uma fonte externa de informação para sua compreensão. Mapear o significado real dos campos fará dos dados mais relevantes e de compreensão mais fácil, além de ampliar as possibilidades de conexão da presente ontologia com outras ontologias existentes. A normalização dos dados e da ontologia e sua disponibilização em plataformas de dados abertos e conectados como a *WikiData*, é o trabalho futuro mais ambicioso desta pesquisa.

Para que tal objetivo seja alcançado há antes a necessidade de otimização da ontologia operacional e de sua prova de funcionamento com outras bases de dados - de modo que questões para além das Questões de Competência estabelecidas nesta pesquisa sejam respondidas. Outra progressão bastante natural é a inclusão de dados sobre cursos das áreas de ciências, engenharias, tecnologias (para além de Computação e TICs) e matemática. Tal expansão exige embasamento teórico sobre o que se define como STEM - termo amplo e aberto a diversas interpretações - e quais áreas do conhecimento se localizam dentro dele.

Para a otimização da ontologia elaborada, sugerimos a revisão da nomenclatura - especialmente dos identificadores (<id_course>), IRIs e termos de geolocalização. São

pontos cruciais nesse processo a verificação de como ontologias mais maduras - e que já estão no padrão 5 Estrelas de Dados Conectados (Figura 4) - nomeiam entidades com as quais também lidamos, assim como o alinhamento aos padrões da W3C.

Outro trabalho futuro importante é a incorporação dos dados do Censo do Ensino Superior do Inep relativo a outros anos - trabalho que há de exigir reestruturação da ontologia para que esta lide com a temporalidade dos dados. Mais um fator a ser endereçado são as entradas relativas a ensino remoto, especialmente as instâncias cujos campos relativos a localização são vazios.

O presente trabalho traz uma contribuição tangível e significativa ao projeto e ao objetivo de se compreender melhor o contexto atual de mulheres em carreiras de STEM. O faz ao estabelecer uma metadologia para elaboração de ontologias adequada ao ELLAS, instanciá-la no contexto da computação no Brasil e por triplificar os dados do Inep de CSV para RDF. Uma das questões de competência respondidas explicita que apenas 14,2% das pessoas formadas em cursos de Sistemas de Informação no país em 2021 eram do sexo feminino - trazendo objetividade a uma percepção da área que já era bastante sólida. A materialidade que dados assim têm possibilita argumentos mais embasados para que, por exemplo, programas e políticas públicas com foco em paridade de gênero na TI sejam desenvolvidas. São esses 14,2% que explicitam a relevância deste trabalho.

Referências

ABDELGHANY, A.; DARWISH, N.; HEFNI, H. An agile methodology for ontology development. **Int. j. intell. eng. syst.**, The Intelligent Networks and Systems Society, v. 12, n. 2, p. 170–181, 2019. Citado 2 vezes nas páginas 24 e 25.

ARZBERGER, P. W. et al. An international framework to promote access to data. **Science**, v. 303, p. 1777 – 1778, 2004. Citado na página 23.

BERNERS-LEE, T. **Linked data**. W3C, 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData>. Citado 3 vezes nas páginas 23, 24 e 25.

ENGLISH, L. D. STEM education k-12: perspectives on integration. **Int. j. STEM educ.**, Springer Science and Business Media LLC, v. 3, n. 1, dez. 2016. Citado na página 19.

FALBO, R. D. A. Sabio: Systematic approach for building ontologies. In: . Rio de Janeiro: ONTO-COM-ODISE 2014, 2014. Disponível em: <http://ceur-ws.org/Vol-1301/ontocomodise2014_2.pdf>. Citado 3 vezes nas páginas 25, 26 e 27.

GIL, A. C. **Como elaborar projetos de pesquisa**. 7th. ed. Barueri: Editora Atlas. ISBN 9786559771646. Citado na página 34.

GRANT, M. J.; BOOTH, A. A typology of reviews: an analysis of 14 review types and associated methodologies. **Health Information &amp Libraries Journal**, Wiley, v. 26, n. 2, p. 91–108, maio 2009. Disponível em: <https://doi.org/10.1111/j.1471-1842.2009.00848.x>. Citado na página 28.

GRUNINGER, M.; FOX, M. S. Methodology for the design and evaluation of ontologies. In: . [S.l.: s.n.], 1995. Citado na página 35.

GUARINO, N. **Formal ontology in information systems : proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy**. Amsterdam Washington, DC Tokyo: IOS Press Omsha, 1998. ISBN 978-90-5199-399-8. Citado na página 24.

ISOTANI, S.; BITTENCOURT, I. I. **Dados abertos conectados**. Sao Paulo: Novatec, 2015. ISBN 9788575224496. Citado 4 vezes nas páginas 15, 16, 21 e 22.

KOONCE, D. A. et al. What is stem? 2011 ASEE Annual Conference and Exposition, Vancouver, Canada, 2011. Citado na página 19.

MACIEL, C.; BIM, S. A.; FIGUEIREDO, K. S. Digital girls program – disseminating computer science to girls in brazil. **GE '18: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering**, Gothenburg, Sweden, 2018. Citado na página 15.

NADELSON, L. S.; SEIFERT, A. L. Integrated stem defined: Contexts, challenges, and the future. **The Journal of Educational Research**, Taylor Francis and Routledge, v. 110, n. 3, p. 221–223, mar. 2017. Citado 2 vezes nas páginas 19 e 20.

NOY, N. F.; MCGUINNESS, D. L. **Ontology Development 101: A Guide to Creating Your First Ontology**. [S.l.], 2001. Citado na página 25.

OSTENDORFF, M.; BLUME, T.; OSTENDORFF, S. Towards an open platform for legal information. In: **Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020**. ACM, 2020. Disponível em: <https://doi.org/10.1145/3383583.3398616>. Citado 2 vezes nas páginas 29 e 30.

SHIRAI, S. S. et al. Identifying ingredient substitutions using a knowledge graph of food. **Frontiers in Artificial Intelligence**, Frontiers Media SA, v. 3, jan. 2021. Disponível em: <https://doi.org/10.3389/frai.2020.621766>. Citado na página 29.

SILVA, D. da; ZIVIANI, A.; PORTO, F. Aprendizado de máquina e inferência em grafos de conhecimento. In: **Tópicos em Gerenciamento de Dados e Informações: Minicursos do SBBD 2019**. SBC, 2019. p. 93–122. Disponível em: <https://doi.org/10.5753/sbc.6251.1.4>. Citado 2 vezes nas páginas 23 e 24.

SUÁREZ-FIGUEROA, M. C.; GÓMEZ-PÉREZ, A.; FERNÁNDEZ-LÓPEZ, M. The NeOn methodology for ontology engineering. In: **Ontology Engineering in a Networked World**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 9–34. Citado 3 vezes nas páginas 25, 26 e 27.

TANNENBAUM, C. et al. Sex and gender analysis improves science and engineering. **Nature**, v. 575, p. 137—-146, 2019. Citado na página 27.

THORNTON, K.; SEALS-NUTT, K. Science stories: Using iiif and wikidata to create a linked-data application. In: **SEMWEB**. [S.l.: s.n.], 2018. Citado na página 31.

THORNTON, K. et al. Linking women editors of periodicals to the wikidata knowledge graph. **Semantic Web journal**, v. 13, 2021. Citado 4 vezes nas páginas 20, 22, 30 e 31.

Apêndices

APÊNDICE A – Objetivos do Projeto Latin American Open Data for Gender Equality Policies Focusing on Leadership in STEM

O objetivo principal do projeto é o desenvolvimento de uma plaforma de dados abertos e conectados que possibilite intervenções visando a redução da lacuna de gênero em STEM, para promover a discussão pública destinada a aumentar o número de mulheres líderes em universidades, indústrias e instituições públicas, e para promover o desenvolvimento de aplicativos móveis e web baseados em dados abertos, com vistas a aumentar a consciência da importância das mulheres em STEM.

Os objetivos específicos mapeados para que, ao longos dos 3 anos de projeto, seja alcançado seu objetivo principal são:

1. Mapear os fatores, atores e políticas que influenciam o desenvolvimento da carreira das mulheres em STEM, coletar dados relacionados e analisar esses dados.
2. Construir e implantar uma plataforma de dados abertos que integre dados primários e secundários sobre o crescimento da carreira de mulheres em STEM;
3. Promover o uso de dados abertos sobre mulheres na liderança de STEM para aumentar a consciência pública sobre as questões de gênero na área.
4. Fornecer recomendações aos formuladores de políticas na América Latina para aumentar a representação feminina em STEM com foco na igualdade e diversidade de gênero.

Para que tais objetivos sejam alcançados, estabeleceram-se diversas atividades que foram divididas entre as 3 fases do projeto, sendo que cada fase representa um ano. O presente trabalho tange as atividades iniciais - fase e ano 1 - do projeto.

APÊNDICE B – Dicionário de Dados ELLAS

Nome Inep	Nome ELLAS Ontology	Descrição	Mapeamento	Categoria
n/a	id_course	Variável criada a partir da concatenação das colunas course, enrolled_masc, hei_acronym e city, é o identificador único de cada instância na ontologia	Classe	
NO_CINE_AREA_DETALHADA	specific_area	Nome da área detalhada, conforme adaptação da Classificação Internacional Normalizada da Educação Cine/Unesco	Classe	
NO_CINE_AREA_GERAL	general_area	Nome da área geral, conforme adaptação da Classificação Internacional Normalizada da Educação Cine/Unesco	Classe	
NO_CINE_ROTULO	course_cine	Nome do curso, conforme adaptação da Classificação Internacional Normalizada da Educação Cine/Unesco	Classe	
NO_CURSO	course	Nome do Curso	Classe	
NO_IES	hei	Nome da Instituição de Ensino Superior (IES)	Classe	
NO_MUNICIPIO_IES	city	Nome do Município da sede administrativa ou reitoria da IES	Classe	
NO_REGIAO_IES	region	Nome da região geográfica da sede administrativa ou reitoria da IES	Classe	
NO_UF_IES	state	Nome da Unidade da Federação da sede administrativa ou reitoria da IES	Classe	
IN_CAPITAL_IES	is_capital	Informa se a sede administrativa ou reitoria da IES está localizada na capital da Unidade da Federação	Propriedade de dados	0. Não 1. Sim
IN_GRATUITO	is_free	Informa se o curso é gratuito	Propriedade de dados	0. Não 1. Sim
QT_CONC_FEM	seniors_fem	Quantidade de concluintes do sexo feminino	Propriedade de dados	
QT_CONC_MASC	seniors_masc	Quantidade de concluintes do sexo masculino	Propriedade de dados	
QT_ING_FEM	freshman_fem	Quantidade de ingressantes do sexo feminino	Propriedade de dados	
QT_ING_MASC	freshman_masc	Quantidade de ingressantes do sexo masculino	Propriedade de dados	
QT_MAT_FEM	enrolled_fem	Quantidade de matrículas do sexo feminino	Propriedade de dados	
QT_MAT_MASC	enrolled_masc	Quantidade de matrículas do sexo masculino	Propriedade de dados	
SG_IES	hei_acronym	Sigla da IES	Propriedade de dados	
SG_UF_IES	state_acronym	Sigla da Unidade da Federação da sede administrativa ou reitoria da IES	Propriedade de dados	

TP_CATEGORIA_ADMINISTRATIVA	academic_administration	Tipo de Categoria Administrativa da IES (Pública Federa , Privada sem fins lucrativos...)	Propriedade de dados	1. Pública Federal 2. Pública Estadual 3. Pública Municipal 4. Privada com fins lucrativos 5. Privada sem fins lucrativos 6. Privada - Particular em sentido estrito 7. Especial 8. Privada comunitária 9. Privada confessional
TP_GRAU_ACADEMICO	academic_degree	Tipo do grau acadêmico conferido ao ao aluno pela conclusão dos requisitos exigidos pelo curso	Propriedade de dados	1. Bacharelado 2. Licenciatura 3. Tecnológico 4. Bacharelado e Licenciatura (.) Não aplicável
TP_MODALIDADE_ENSINO	instructional_modality	Tipo de modalidade de ensino do curso	Propriedade de dados	1. Presencial 2. Curso a distância
TP_ORGANIZACAO_ACADEMICA	academic_organization	Tipo de Organização Acadêmica da IES	Propriedade de dados	1. Universidade 2. Centro Universitário 3. Faculdade 4. Instituto Federal de Educação, Ciência e Tecnologia 5. Centro Federal de Educação Tecnológica
TP_REDE	funding_category	Rede de Ensino	Propriedade de dados	1. Pública 2. Privada

APÊNDICE C – Consultas em SPARQL para responder às Questões de Competência (QCs)

Algoritmo 1: (QC 1) Quantas mulheres ingressaram em cursos de computação no Brasil em 2021?

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ellas:
 <http://www.semanticweb.org/vanessa/ontologies/2022/11/ellas_ontology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT (SUM(xsd:integer(?freshman_fem_2021)) As
 ?Total_de_ingressantes)

WHERE {

?id_course rdf:type ellas:Id_course .
?id_course ellas:freshman_fem_2021 ?freshman_fem_2021 .

#Resultado: 50704

}
```

Algoritmo 2: (QC 1.1) Quantas mulheres ingressaram em cursos de computação em 2021 em Curitiba?

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ellas:
 <http://www.semanticweb.org/vanessa/ontologies/2022/11/ellas_ontology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT (SUM(xsd:integer(?freshman_fem_2021)) As
 ?Total_de_ingressantes_Curitiba)

WHERE {

?id_course rdf:type ellas:Id_course.
?id_course ellas:freshman_fem_2021 ?freshman_fem_2021 .
?city rdf:type ellas:City.
?id_course ellas:id_city ?city.
?city rdfs:label "CURITIBA"

#Resultado: 1391

}
```

Algoritmo 3: (QC 2) Quantas mulheres concluíram bacharelados em cursos de computação em universidades públicas no Brasil em 2021?

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ellas:
 <http://www.semanticweb.org/vanessa/ontologies/2022/11/ellas_ontology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT (SUM(xsd:integer(?seniors_fem_2021)) As
 ?Concluintes_Universidades_Publicas)

WHERE {

?id_course rdf:type ellas:Id_course .
?id_course ellas:seniors_fem_2021 ?seniors_fem_2021 .
?id_course ellas:academic_degree "1". #Bacharelado .

?hei rdf:type ellas:Hei .
?id_course ellas:id_has_hei ?hei .
?hei ellas:funding_category "1"#Universidade pública

#Resultado: 737

}
```

Algoritmo 4: (QC 3) Entre as pessoas matriculadas dos cursos de Sistemas de Informação em capitais do país em 2021, qual a porcentagem de mulheres?

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ellas:
 <http://www.semanticweb.org/vanessa/ontologies/2022/11/ellas_ontology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ( SUM(xsd:integer(?seniors_fem_2021)) * 100 /
 (SUM(xsd:integer(?seniors_fem_2021)) +
 SUM(xsd:integer(?seniors_masc_2021))) As ?porcentagem )

WHERE {

?id_course rdf:type ellas:Id_course .
?id_course ellas:seniors_fem_2021 ?seniors_fem_2021 .
?id_course ellas:seniors_masc_2021 ?seniors_masc_2021 .

?course rdf:type ellas:Course .
?id_course ellas:id_has_course ?course .
?course rdfs:label "SISTEMAS DE INFORMACAO".

?city rdf:type ellas:City .
?id_course ellas:id_city ?city .
?city ellas:is_capital "1"# É capital

#Resultado: 14.198%

}
```

Algoritmo 5: (QC 4) Quantas mulheres estavam matriculadas em Institutos Federais de Educação, Ciência e Tecnologia na área da computação no Nordeste em 2021?

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ellas:
 <http://www.semanticweb.org/vanessa/ontologies/2022/11/ellas_ontology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT (SUM(xsd:integer(?enrolled_fem_2021)) As
 ?Alunas_IFs_Nordeste)

WHERE {
?id_course rdf:type ellas:Id_course .
?id_course ellas:enrolled_fem_2021 ?enrolled_fem_2021 .

?regiom rdf:type ellas:Region .
?id_course ellas:id_region ?region.
?region rdfs:label "NORDESTE".

?hei rdf:type ellas:Hei .
?id_course ellas:id_has_hei ?hei .
?hei ellas:academic_organization "4". # Instituto Federal de Educação,
 Ciência e Tecnologia

#Resultado: 1662

}
```