Construção de grafos de conhecimento baseados em dados não-estruturados com envolvimento de especialistas de domínio

Pedro Henrique Stolarski Auceli[1], Rita Cristina Galarraga Berardi[1]

[1]*Federal University of Technology - Paraná*

Abstract

The development of a Knowledge Graph (KG) is no easy task since there's a need to understand the domain in order to correctly define how it works. The role of ontologies in Knowledge Graphs are related to the structure and definition of the data from a domain. The involvement of domain experts in the development of ontologies is a topic that has and is being researched, but the involvement of these experts in the process of developing a Knowledge Graph is still relatively underexplored, especially when the KG is built by using non-structured data. This paper presents a methodology to construct Knowledge Graphs with non-structured data involving domain experts in the process of development. In order to define this methodology a case study was conducted during a projet called ELLAS which has a group of domain experts from the gender equality domain, with steps that were adapted from the literature. The insights from this case study contributed to the iterative development of the methodology.

Keywords

Knowledge graph, Domain experts, Knowledge graph engineering

1. Introdução

Grafos de conhecimento (GC) são redes semânticas que integram diversas fontes heterogêneas de dados para representar o conhecimento de um certo domínio [1]. As ontologias são utilizadas nestes grafos para a definição semântica dos conceitos e para a estruturação dos dados, então com a utilização dessa os dados são inseridos em nós e arestas do grafo para representar a realidade do domínio [2][3]. O termo ficou famoso a partir da publicação da Google sobre o seu GC, que inspirou diversas outras empresas como Amazon, Airbnb e Facebook a investirem nesta tecnologia. Os GC se destacaram por organizar, integrar, permitir a consulta de dados de bases vastas que podem ser aplicadas a diversos domínios [2].

A construção de um GC necessita de dados, sendo que esses podem ser dos mais diversos tipos, como: estruturados, não-estruturados, semi-estruturados, homogêneos ou heterogêneos. Desses tipos os não-estruturados e semi-estruturados são os mais complexos de se lidar pois além do processo de descoberta dos dados é necessário o processo de extração de conhecimento para determinar a estrutura deles[1]. Os dados não-estruturados, no contexto de banco de dados, são dados que não podem ser inseridos em colunas e linhas mais comumente encontrados em textos, vídeos, áudios e imagens [4].

Como as ontologias são necessárias para a construção de GC a obtenção do conhecimento sobre o domínio que se deseja modelar é necessária, tarefa que não é considerada trivial [5]. O passo para obtenção do conhecimento do domínio é chamado de análise de domínio, o qual pode ser feito de várias maneiras como a extração através da literatura, extração do conhecimento dos especialistas do domínio ou a mistura desses dois métodos [6].

Autores defendem o envolvimento de especialista do domínio no processo de engenharia de uma ontologia, devido ao seu conhecimento prévio sobre o domínio [7]. A inserção de especialistas de domínio enriquece a estrutura lógica de uma ontologia devido ao conhecimento prévio que os especialistas têm sobre o contexto, o que evita definições incorretas que podem prejudicar o modelo final [8].

Proceedings of the 17th Seminar on Ontology Research in Brazil (ONTOBRAS 2024) and 8th Doctoral and Masters Consortium on Ontologies (WTDO 2024), Vitória, Brazil, October 07-10, 2024.

pedroauceli@alunos.utfpr.edu.br (P. H. S. Auceli); ritaberardi@utfpr.edu.br (R. C. G. Berardi)

0009-0000-1903-7960 (P. H. S. Auceli); 0000-0002-0281-8952 (R. C. G. Berardi)

Logo o problema desta pesquisa é explorar os desafios encontrados no devido envolvimento de especialistas do domínio na criação de um GC através de dados não-estruturados provenientes de texto. Onde existem diversos desafios como: como criar a ontologia utilizada no GC, como preparar os dados não-estruturados para inserção no GC e como envolver especialistas neste processo. A hipótese deste trabalho é que, graças ao envolvimento de especialistas do domínio no processo de desenvolvimento, tanto o processo de curadoria dos dados não-estruturados, quanto da construção do grafo, será aprimorada devido ao conhecimento prévio dos especialistas.

Logo, o objetivo deste trabalho é a criação de um processo para envolver especialistas de domínio na construção de GC baseados em dados não-estruturados provenientes de textos. Esta é uma pesquisa exploratória que é desenvolvida através da condução de um estudo de caso com especialistas do domínio de um projeto que desenvolve uma plataforma de grafos de conhecimento para questões de equidade de gênero, chamado ELLAS (Equality in Leadership for Latin American STEM) envolvendo especialistas desse domínio do Brasil, Bolívia e Peru com heterogeneidade de língua, conceito e formações acadêmicas, englobando Computação, Psicologia, Educação e Negócios.

2. Engenharia de grafos de conhecimento

Os passos para a criação de um GC são variados e diferentes autores definem passos distintos que devem ser seguidos no processo de construção. Os seguintes passos são um compilado de passos realizado pelo autor do presente trabalho entre [3] e [1], sendo que isto foi realizado para melhor englobar o processo de desenvolvimento.

O primeiro passo é a aquisição dos dados e do processamento desses, onde as tarefas pertinentes são a obtenção dos dados e processamento deles, como redução de granularidade ou remoção de ruído. O próximo passo é a extração do conhecimento, onde a análise do domínio para o desenvolvimento da ontologia que será utilizada no GC deve ser executada. Porém como nos GC os dados são indispensáveis é necessário fazer a análise da estrutura dos dados provenientes do primeiro passo, para assim desenvolver uma ontologia que reflete o conhecimento do domínio e dos dados disponíveis.

O terceiro passo é a definição dos metadados, onde, novamente, são necessárias ontologias. Esses dados ajudam a enriquecer o grafo pois trazem informações sobre os dados em si como por exemplo: a proveniência dos dados, metadados sobre a estrutura, informação temporal, reportes sobre a qualidade e logs de processamento. O quarto passo é a fusão de entidades. Como os dados podem ser provenientes de diversas fontes, de diversos domínios, não é possível assegurar que termos com o mesmo significado, ou a mesma semântica, terão o mesmo valor, como por exemplo em uma base Brasil pode ter o valor **Brasil** enquanto em outra pode ter o valor **Brazil**, ambos termos escritos diferente mas com o mesmo sentido de país. Logo é necessária a fusão de termos diferentes que têm a mesma semântica, isto é geralmente realizado através de links owl:sameAs entre os conceitos.

Por fim, é necessário o enriquecimento do GC. Como a base da estrutura dos GC é ditada por uma ontologia, as regras lógicas definidas podem ser utilizadas para incluir novos dados dentro do grafo. Isto é mais comumente utilizado em treinamentos de *machine learning*, pois a aquisição de novos dados, de novas relações e o enriquecimento de dados específicos do domínio aumentam a representatividade do grafo o que pode auxiliar no treinamento dos algoritmos.

3. Metodologia da pesquisa

A metodologia desta pesquisa foi criada com o intuito de entender e aprender com o processo de desenvolvimento de um GC sobre dados não-estruturados provenientes de texto com o envolvimento de especialistas do domínio.

3.1. Levantamento do estado da arte sobre análise de domínio

Como existe uma brecha no tópico de envolvimento de especialistas do domínio na construção de um GC foi necessário o aprofundamento de um dos passos essenciais na criação de ontologias, a análise de domínio. Sendo que este foi escolhido por ser um passo com um grande envolvimento de especialistas de domínio. Logo, um mapeamento sistemático foi realizado de acordo com o protocolo de [9] . Com o foco nos métodos para analisar um domínio para a construção de ontologias, sejam eles com ou sem o envolvimento de especialistas de domínio no processo.

A pesquisa dos artigos deste mapeamento foi realizada em 3 plataformas: Science Direct, Google Scholar e ACM. Além disso, string de buscas, e critérios de exclusão foram definidos, como a presença de palavras chaves no resumo, e a criação de perguntas que deveriam ser respondidas através da leitura. No fim foram escolhidos 14 artigos para a leitura completa. Sendo que foram encontradas 3 categorias de método para análise de domínio: análise da literatura, extração de conhecimento de especialistas do domínio e mistas ou outras. A análise a partir da literatura é feita através do estudo de textos técnicos, manuais, artigos científicos e possivelmente de diagramas UML. A análise de domínio feita através da extração de conhecimento de especialistas do domínio, ou da inserção do engenheiro do conhecimento na realidade dos especialistas, se baseia em obter o conhecimento através de alguma técnica. Essas técnicas podem ser surveys, um modelo baseado no domínio, entrevistas ou até uma técnica focada em dados como a utilização de planilhas. A análise de domínio mista é feita através da mistura dos 2 passos anteriores. Estes métodos dividem a análise em 2 passos, a análise da leitura é feita antes para obter o conhecimento geral do domínio, e com esse conhecimento criar uma técnica para extração do conhecimento dos especialistas do domínio através das técnicas de survey, entrevistas e modelos baseados no domínio.

3.2. Definição de um método para criação de GC

Como o processo de construção de um GC é complexo e exige diversas competências, foi necessário a definição de um método que pudesse ser aplicado na realidade do estudo de caso.

3.2.1. Processo de treinamento

O primeiro passo definido foi o da construção e aplicação de um treinamento sobre ontologias e GC. Este treinamento deve ser ofertado presencialmente e dividido entre apresentações e dinâmicas. Onde as apresentações devem contextualizar e explicar os conceitos de ontologia e GC, enquanto que as dinâmicas devem focar na estrutura de GC, modelagem de ontologias e suas regras lógicas, sendo que os conceitos utilizados na dinâmica devem ser de um domínio que os especialistas tenham conhecimento. Também deve se evitar um treinamento de ferramentas, pois autores evidenciaram as dificuldades na utilização de ferramentas para a modelagem de um domínio e como treinamentos extensos são necessários para a utilização dessas [10]. Logo, a dinâmica deve ser criada utilizando conceitos de aprendizado cinestésico, uma vez que este método vem sendo aplicado em aprendizados de computação, e que vem trazendo resultados promissores no engajamento dos membros [11].

3.2.2. Metodologia para análise de domínio e estruturação de dados

Como são necessários dados e uma estrutura ontológica para a construção de um GC, foi necessário criar um método para a análise do domínio para a obtenção dos termos e relacionamentos da ontologia e para a busca de dados relevantes ao domínio. Para o caso deste trabalho a técnica de análise de domínio escolhida foi a de análise através da literatura devido a necessidade do processo de curadoria dos dados não-estruturados, onde uma revisão sistemática sobre o tópico deve ser realizada para a obtenção dos conceitos relacionados ao domínio. Esta revisão deve ser realizada pelos próprios especialistas do domínio, pois uma das hipóteses deste trabalho é que devido ao conhecimento do domínio por parte dos especialistas a escolha pelos termos importantes seja mais próxima a realidade do domínio. Os termos encontrados devem ser inseridos em uma planilha, onde os conceitos e os relacionamentos devem ser

transformados em colunas, sendo que estes devem estar presentes em um dicionário de dados com as suas respectivas definições. Também devem ser criadas questões de competência de acordo com a estrutura definida. E por fim os dados não-estruturados encontrados através da revisão sistemática devem ser adicionados nas colunas criadas.

3.2.3. Comunicação entre a equipe

Como a comunicação entre os especialistas do domínio e do engenheiro de conhecimento é essencial durante o processo de desenvolvimento do grafo, devem ser disponibilizadas diversas formas de comunicação, sejam elas síncronas ou assíncronas. Para retirada de dúvidas específicas documentos onde dúvidas podem ser escritas ou chat virtual são formas assíncronas, caso seja necessário também podem ser realizadas reuniões virtuais como forma síncrona de resolução de dúvidas. Como é possível que dúvidas iguais surjam em diferentes membros a criação de uma wiki contendo os passos do desenvolvimento e as dúvidas frequentes, auxilia na redução de interações entre especialistas do domínio e engenheiro de conhecimento, além de agilizar a resolução do problema. Para retirada de dúvidas no processo de desenvolvimento da ontologia e do GC, formas assíncronas foram consideradas ineficientes devido à complexidade das questões pertinentes ao domínio, logo reuniões virtuais com os membros responsáveis pelas definições devem ser o padrão.

3.2.4. Método de validação e procura de novos vocabulários

Como este trabalho ainda está em andamento este passo ainda não foi finalizado. Até o momento da escrita foi decidido que serão utilizados formulários online contendo breves histórias sobre os conceitos definidos na ontologia para validar a definição dada pelos especialistas do domínio.

3.3. Estudo de caso

A escolha pelo método de estudo de caso foi motivada pela presença de um projeto de pesquisa que apresentava um problema prático relevante e passível de ser investigado através de uma abordagem científica. A metodologia de estudo de caso se revelou adequada devido à sua capacidade de oferecer uma análise profunda e contextualizada do problema em questão. Esta abordagem não apenas permite uma exploração minuciosa do problema, como também proporciona insights valiosos para a aplicação prática da metodologia científica. [12]. O estudo de caso foi realizado em conjunto dos especialistas de domínio deste projeto, que tem como foco mapear as iniciativas e políticas que promovem o aumento da presença de mulheres nas áreas STEM *(Science, Technology, Engineering and Mathematics)* e os fatores que influenciam na decisão de escolha e na retenção de profissionais dessas áreas.

Neste estudo os especialistas foram envolvidos no processo de desenvolvimento de um GC sobre o domínio da equidade de gênero em áreas STEM, sendo que o método utilizado foi uma adaptação de "Ontology 101" [7] para a ontologia, e dos passos apresentados na seção 2 para o GC. Sendo que os aprendizados obtidos através da realização dos passos de construção do GC foram utilizados no desenvolvimento do processo proposto. O pipeline com todos os passos deste estudo de caso pode ser visto em mais detalhes nos trabalhos [13][14]. Os dados considerados para gerar esse GC são relacionados ao levantamento de políticas existentes, iniciativas realizadas para aumentar a presença de mulheres nessas áreas, fatores que podem influenciar a escolha e permanência de mulheres nessa área. O enfoque em criar os GC da rede ELLAS é para integrar os dados, homogeneizando o vocabulário e a compreensão de conceitos deste domínio para analisar a situação da América Latina como um todo e assim auxiliar os tomadores de decisão ao tomar ações nesta direção.

Este contexto é motivacional para essa pesquisa porque possui características não comumente encontradas na literatura como diversidade de língua, cultura e principalmente as fontes em que os dados se originam. É bastante comum que esse tipo de dado não esteja disponível estruturado, e sim esteja presente em diversas fontes como artigos científicos, reports, redes sociais e muitas vezes as técnicas computacionais não dão conta de identificar quais conceitos são importantes para discutir essa problemática, então é necessária a ação humana e o envolvimento direto desses pesquisadores na

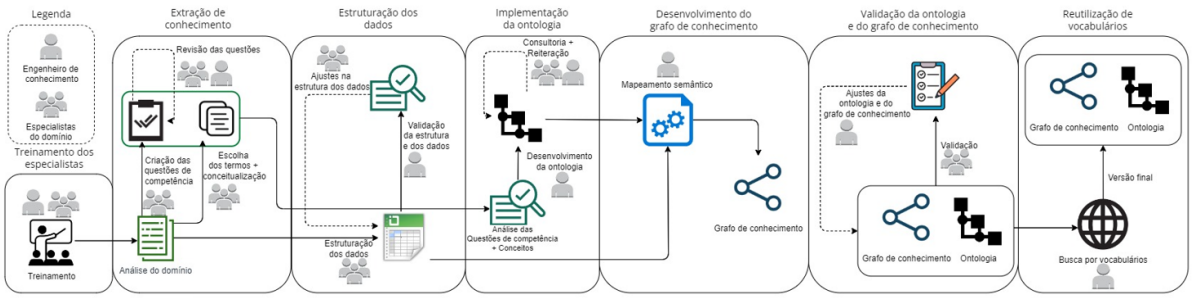

Figure 1: Processo proposto

estruturação dos dados. Esses dados não estruturados exigem uma presença ativa de especialistas de domínio na coleta, na estruturação e conceituação do modelo ontológico.

4. Processo proposto

Neste processo foram definidos 7 passos que deverão ser seguidos para a obtenção de um GC no final com a devida inserção de especialistas de domínio. Esses passos podem ser visualizados na Figure 1. Os 7 passos definidos foram:

1. **Treinamento dos especialistas:** um treinamento deve ser criado e aplicado nos especialistas do domínio sem enfoque em ferramentas de construção de ontologias como Protégé. Recomenda-se a aplicação de um treinamento com base no aprendizado cinestésico com exemplos do domínio escolhido. Também será necessária a criação de um treinamento sobre questões de competência, demonstrando como elas devem ser criadas a partir de dados coletados e como validar elas através de uma matriz de rastreabilidade, é importante que esse treinamento seja realizado ao decorrer do segundo passo. Como os especialistas podem não ter conhecimento sobre estruturação de dados é necessário a criação de um treinamento sobre conceitos básicos, como enriquecimento de dados, desaglomeração de colunas e adição de novos dados de acordo com conhecimento prévio, sendo que este deverá ser aplicado ao decorrer do terceiro passo.
2. **Extração do conhecimento:** especialistas de domínio devem extrair o conhecimento através da análise do domínio e devem escolher os termos importantes para a definição do domínio e conceituá-los, além disso também devem criar questões de competência com os conceitos escolhidos. Por fim, o engenheiro do conhecimento deve validar os termos escolhidos e as questões de competência criadas.
3. **Estruturação dos dados:** os especialistas devem extrair os dados não estruturados das fontes e estruturar esses em uma planilha de acordo com os termos escolhidos. Após a estruturação o engenheiro do conhecimento deve validar os dados, a estrutura da planilha e o dicionário de dados.
4. **Implementação da ontologia:** uma ontologia deve ser desenvolvida a partir dos termos escolhidos e das questões de competência criadas. É essencial o envolvimento dos especialistas neste processo para retirada de dúvidas, devido aos termos escolhidos e as questões de competência terem sido definidas por eles. É importante que a linguagem utilizada entre o engenheiro do conhecimento e dos especialistas do domínio seja simples, uma vez que a utilização de termos técnicos relacionados à ontologia dificulta o entendimento. Logo, é o trabalho do engenheiro traduzir as regras e relacionamentos criados em frases relacionadas ao domínio.
5. **Desenvolvimento do GC:** um GC deve ser criado com base na estrutura da ontologia criada e com os dados fornecidos nas planilhas. O engenheiro do conhecimento deve analisar e escolher as ferramentas que melhor se adequem a sua realidade, será necessário escolher uma para transformar

os dados contidos nas planilhas para a estrutura definida na ontologia e uma triplestore para o armazenamento e consulta dos dados.

6. **Validação da ontologia e do GC:** um método de validação da ontologia e do GC deve ser criado e aplicado juntamente dos especialistas do domínio. A validação da ontologia deve ser feita para validar os conceitos, classes e relacionamentos criados pelo engenheiro do conhecimento. Para o GC devem ser criadas queries para cada questão de competência criada pelos especialistas de domínio.
7. **Reutilização de vocabulários:** o engenheiro do conhecimento deverá analisar a versão final da ontologia e procurar por novos vocabulários que possam definir os conceitos presentes na ontologia, assim aumentando a representatividade da ontologia e facilitando a conexão. Caso sejam feitas mudanças o grafo deverá ser alterado para seguir a nova estrutura.

5. Discussão sobre os resultados

Apesar do estudo de caso não ter sido finalizado, alguns resultados já foram obtidos. Uma primeira versão da ontologia foi disponibilizada para os especialistas do domínio contendo 113 axiomas, 37 classes, 10 propriedades de objeto e 36 propriedades de dados. Essa estrutura foi utilizada para definir os conceitos e relacionamentos entre políticas públicas, iniciativas e fatores que influenciam a presença de mulheres em áreas STEM.

Até o momento da escrita foram criadas 38 questões de competência sobre os 3 tópicos escolhidos, sendo que todas foram possíveis de responder. Alguns exemplos das questões são:Quais tipos de políticas/processos/práticas de gênero existem na América Latina?;Quais iniciativas foram feitas para um tipo específico de pessoas com vulnerabilidades ?;Quais são os tipos de impacto de um fator contextual específico em instituições latino americanas?

Vale ressaltar que algumas questões de competência demonstram as vantagens da integração dos dados através de um GC, onde é possível verificar a conexão entre os dados de acordo com relacionamentos em comum, como é o caso da seguinte questão de competência: Quais políticas, iniciativas e fatores tem o mesmo público alvo? Onde foi visto que o o público mais atingido foi de alunos de graduação.

Durante a execução do estudo de caso foram necessárias várias iterações da metodologia da pesquisa até se chegar em um resultado satisfatório. As maiores dificuldades encontradas nesse processo foi a dificuldade de se encontrar formas de ensino lúdicas, que auxiliassem os especialistas de domínio com as suas dúvidas e aumentar o engajamento de membros nas atividades assíncronas. Também foi visualizada a grande dependência dos especialistas do domínio com o engenheiro de conhecimento, que em momentos do desenvolvimento sobrecarregaram o engenheiro prejudicando o processo proposto. Por fim, apesar das escolhas das ferramentas de comunicação terem sido feitas para englobar o diverso público de especialistas de domínio ainda assim foram encontradas dificuldades como: membros não sendo inseridos nos grupos corretos, membros não visualizando notificações, membros escolhendo formas mais informais para trocas de informação.

6. Considerações finais

Neste trabalho foi apresentado um processo para a criação de grafos de conhecimento sobre dados não-estruturados provenientes de texto com o envolvimento de especialistas de domínio no processo de desenvolvimento. Onde os passos definidos na metodologia foram criados a partir dos aprendizados de um estudo de caso realizado com especialistas do domínio de equidade de gênero. Espera-se que esse método possa ser replicado em diferentes domínios, desde que as restrições com relação aos tipos de dados sejam respeitadas. Como trabalhos futuros ainda é necessário definir o método de validação e busca por vocabulários que deve ser aplicado ao estudo de caso, que posteriormente deverá ser adicionado como passo no processo proposto. O GC será incrementado com novos dados e conceitos, além das mudanças de vocabulário que será resultado do passo de busca por vocabulários, logo o processo proposto ainda sofrerá mudanças ao decorrer desta pesquisa.

References

[1] D. Fensel, U. Şimşek, K. Angele, E. Huaman, E. Kärle, O. Panasiuk, I. Toma, J. Umbrich, A. Wahler, D. Fensel, et al., Introduction: what is a knowledge graph?, Knowledge graphs: Methodology, tools and selected use cases (2020) 1–10.

[2] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, ACM Computing Surveys (Csur) 54 (2021) 1–37.

[3] M. Hofer, D. Obraczka, A. Saeedi, H. Köpcke, E. Rahm, Construction of knowledge graphs: Current state and challenges, Available at SSRN 4605059 (????).

[4] R. Blumberg, S. Atre, The problem with unstructured data, Dm Review 13 (2003) 62.

[5] L. F. Garcia, F. H. Rodrigues, A. G. L. Júnior, R. d. S. A. Kuchle, M. Perrin, M. Abel, What geologists talk about: Towards a frequency-based ontological analysis of petroleum domain terms., in: ONTOBRAS, 2020, pp. 190–203.

[6] R. Smiraglia, Domain analysis for knowledge organization: tools for ontology extraction, Chandos Publishing, 2015.

[7] N. F. Noy, D. L. McGuinness, et al., Ontology development 101: A guide to creating your first ontology, 2001.

[8] E. Norris, J. Hastings, M. M. Marques, A. N. F. Mutlu, S. Zink, S. Michie, Why and how to engage expert stakeholders in ontology development: insights from social and behavioural sciences, Journal of Biomedical Semantics 12 (2021) 1–8.

[9] B. Kitchenham, Procedures for performing systematic reviews, Keele, UK, Keele University 33 (2004) 1–26.

[10] V. Dimitrova, R. Denaux, G. Hart, C. Dolbear, I. Holt, A. G. Cohn, Involving domain experts in authoring owl ontologies, in: The Semantic Web-ISWC 2008: 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings 7, Springer, 2008, pp. 1–16.

[11] R. Berardi, N. Kozievitch, S. A. Bim, P. Auceli, Oficina de banco de dados com aprendizado cinestésico para meninas do ensino médio, in: Anais do Workshop de Informática na Escola, volume 25, 2019, pp. 345–354.

[12] R. Heale, A. Twycross, What is a case study?, 2018.

[13] R. C. G. Berardi, P. H. S. Auceli, C. Maciel, G. Davila, I. R. Guzman, L. Mendes, Ellas: Uma plataforma de dados abertos com foco em lideranças femininas em stem no contexto da américa latina, in: Anais do XVII Women in Information Technology, SBC, 2023, pp. 124–135.

[14] R. C. G. Berardi, P. H. S. Auceli, C. Maciel, R. Fritoli, G. Davila, I. R. Guzman, L. Mendes, Ellas architecture and process: Collecting and curating data on women's presence in stem, Journal on Interactive Systems 15 (2024) 530–540.