Infraestrutura da arquitetura e pipeline de automação da governança de um grafo de conhecimento com fontes de dados estruturadas e não estruturadas: Uma aplicação na plataforma ELLAS

Rodgers Fritoli[1,*,†], Rita Cristina Galarraga Berardi[2,†]

[1]*Projeto Latin American Open Data for Gender Equality Policies Focusing on Leadership in STEM(ELLAS), Brasil*
[2]*Programa de Pós Graduação em Computação Aplicada (PPGCA), Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil*

Abstract

Knowledge graph technology is rapidly evolving, as it enables the integration of different types of data into a semantic context, facilitating complex analyses and innovative discoveries. However, the creation of an initial architecture for these graphs faces challenges such as choosing the right infrastructure to support the volume and variety of data, while ensuring performance and security. In the context of the ELLAS project, which involves data from various sources across Latin America, a systematic literature review was conducted to identify the best infrastructure for governance of these data. The review focused on scalable and flexible solutions suitable for the growth of the graph over time, and explored technologies such as clustering, Big Data, relational and graph databases, and cloud servers. The results contribute to the discussion of appropriate infrastructures for practical applications of knowledge graphs, addressing an existing gap in the literature.

Keyworks

Arquitetura, Graphdb, infraestrutura triplestore, dados conectados

1. Introdução

De acordo com a Gartner, renomada consultoria em tecnologia, a tecnologia de grafos de conhecimento está em desenvolvimento acelerado, pois permite relacionar diferentes tipos de dados em um contexto semântico, permitindo análises complexas e descobertas inovadoras em diversos setores[1] . Existem diferentes formas de armazenar um grafo como notação JSON, modelos de grafo personalizados, bancos de dados de grafos (*Neo4j*, *OrientDB*) ou *triplestores* em RDF (*Resource Description Framework*) ou *OWL - Ontology Web Language* (*Virtuoso*, *GraphDB*)[2]. O RDF permite a utilização de um vocabulário controlado ou ontologias, garantindo consistência semântica na integração de dados com terminologias não padronizadas e com significados diversos. O desenvolvimento de ontologias é eficaz para a integração de diferentes sistemas e fontes de dados, pois permitem realizar consultas e inferências mais avançadas, descobrindo novos conhecimentos a partir dos dados conectados a partir da utilização de *reasoning*[1].

O objetivo deste estudo é determinar a melhor infraestrutura para a governança de um grafo de conhecimento em RDF. Ao definir a infraestrutura inicial, é preciso considerar o contexto específico da aplicação, escolhendo tecnologias que atendam as demandas de armazenamento, processamento, sistema operacional e formas de disponibilização de acessos. A falta de uma infraestrutura ideal pode levar a desempenho lento, escalabilidade limitada, gargalos de processamento, vulnerabilidades de

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.

*Corresponding author.
†These authors contributed equally.
rodfritoli@hotmail.com (R. Fritoli); ritaberardi@utfpr.edu.br (R. C. G. Berardi)
0009-0001-7696-7600 (R. Fritoli); 0000-0002-0281-8952 (R. C. G. Berardi)

[1]https://www.gartner.com/en/articles/understand-and-exploit-gen-ai-with-gartner-s-new-impact-radar
[2]https://graphdb.ontotext.com/documentation

segurança e custos excessivos ou até mesmo custos desnecessários quando da implementação de recursos desproporcionais ao seu uso.

A motivação para esta pesquisa é a construção de uma plataforma de dados abertos conectados que centraliza em um grafo de conhecimento dados dispersos em formatos não estruturados (como artigos acadêmicos em publicações textuais) como também dados estruturados abertos que são trazidos para a plataforma de forma automatizada e com um enfoque específico em 3 línguas diferentes (português, espanhol e inglês). Esta plataforma é da rede ELLAS de pesquisa[3], que é uma rede de pesquisa no contexto da América Latina com a participação de pesquisadores do Brasil, Peru e Bolívia que propõe uma plataforma que coleta, centraliza, integra em um grafo de conhecimento dados abertos sobre variáveis que podem estar relacionadas à baixa presença de mulheres em cursos e carreiras de áreas STEM (Ciência, Tecnologia, Engenharia e Matemática).Para aplicações de dados abertos conectados (ou comumente conhecidos com o termo grafos de conhecimento após o Google mencioná-lo [29]), se consideram inúmeras coletas de dados RDF disponíveis na Web e para processar tais volumes. Embora estudos anteriores indicam que arquiteturas distribuídas [4,5], ou uma arquitetura em cluster [6,7] sejam ideais, este estudo demonstra a aplicabilidade de uma infraestrutura mais simples em um grafo de menor escala e por isso requer um estudo mais dedicado a essas características.A arquitetura completa da plataforma ELLAS foi definida e explicada em [24] onde foram definidas as fontes de dados, as camadas de processamento, entradas e saídas de cada componente, no entanto as definições para a infraestrutura que implementa esta arquitetura está exposta neste artigo assim como toda a metodologia empregada para defini-la.

2. Revisão Sistemática da Literatura

Com o objetivo de identificar as melhores práticas e lacunas de conhecimento sobre infraestrutura de grafos de conhecimento com RDF, conduzimos uma revisão sistemática da literatura. Nesta seção, detalhamos a metodologia empregada, apresentamos os principais resultados e discutimos suas implicações.

2.1. Metodologia da RSL

A RSL seguiu critérios bem definidos, foi utilizado a base de dados *ACM Digital Library*[4] e IEEE[5] . A estratégia de busca adotada foi delineada através da seguinte string: "*architecture* AND *graphdb* OR (*triplestore infrastructure*) OR (*semantic pipeline*) OR (*linked data pipeline*)". Na base ACM Digital Library, retornou o total de 211 artigos com as *strings* pesquisadas, enquanto na IEEE o resultado foi de 38. Após a exclusão de 6 artigos duplicados, foi realizada uma seleção preliminar com base nos títulos, identificando 104 artigos que demonstravam relevância com o tema do estudo. A etapa subsequente envolveu uma análise mais aprofundada, com a leitura dos resumos desses 104 artigos, resultando na seleção criteriosa de 13 publicações para uma leitura integral de seus conteúdos.

2.2. Artigos Resultantes da RSL

Nos artigos [11] e [12] foi indicado como melhor estratégia de armazenamento e gerenciamento dos arquivos RDF na nuvem a utilização de *Hadoop*. Os autores propõem a utilização do modelo de programação *MapReduce* [9] para realizar pesquisas em grandes volumes de dados. Também é sugerido o *Hadoop* no artigo [23], onde foram comparadas duas tecnologias, uma com os dados no *Hadoop* como o banco de dados HBase e outra tecnologia com o banco de dados relacional *MySQL Cluster*. Como infraestrutura são utilizadas nove máquinas comuns para criar o *cluster*. O volume de dados foi de um conjunto entre 38.600 e 80.043.000 de triplas. O artigo [21] apresenta uma infraestrutura que envolve um *cluster* com *Hadoop*. Foi relatado que o *cluster* contém 20 nós sem mencionar especificamente

[3]https://ellas.ufmt.br/pt/inicio/
[4]https://dl.acm.org/
[5]https://ieeexplore.ieee.org/

qual o *hardware* de cada nó. No artigo [13], é implementada uma arquitetura para o armazenamento e disponibilização de dados RDF utilizando o banco de dados *NoSQL Amazon SimpleDB*, e o serviço de armazenamento no *Amazon Simple Storage Service* (S3), para criação de máquinas virtuais é utilizado serviço *Amazon Elastic Compute Cloud* (EC2) e o serviço de fila *Amazon Simple Queue Service* (SQS), possibilitando filas de execução síncronas de consultas. Nos artigos [14] e [15]" ambos não expõem qual infraestrutura foi utilizada para o armazenamento do RDF, somente é detalhado como foi construído um sistema com a utilização de RDF e OWL mas não é mencionado qual foi o servidor ou outra característica da sua infraestrutura. Já no artigo [16] são detalhados os testes realizados com o banco de dados de grafos *Virtuoso*[6] e o *Blazegraph*[7] .Para a configuração do servidor foi utilizado como cloud a Amazon com o serviço *Amazon Elastic Compute Cloud* (EC2) para armazenamento dos dados brutos do *Simple Storage Solutions*(S3). Na pesquisa [17] foi abordada sobre a modelagem e o design do banco de dados para o RDF. É relatado no artigo que o design físico do banco de dados RDF é difícil e não há consenso sobre a melhor abordagem, o desempenho pode ser ruim para diferentes cargas de dados. Na pesquisa foram mencionados testes de criação do banco de dados com RDF-3x 8.9ms, *GStore*, *MonetDB* e *Virtuoso*, embora não tenha sido descrito qual a configuração da infraestrutura de servidores que foi utilizada. No artigo [18] é proposta a criação de um banco de dados RDF relacional baseado em caminhos, funcionando de forma independente das informações dos esquemas RDF. Quanto à volumetria de dados foi utilizado o Gene Ontology[8] que é um consórcio internacional que fornece um vocabulário estruturado e padronizado para descrever funções dos genes, seu volume total é de 45.443 termos. O artigo [19] detalha uma aplicação desenvolvida pelo governo brasileiro para uma infraestrutura de dados abertos utilizando RDF Este trabalho relata um caso em que o governo disponibilizou um grande volume de dados em diversos formatos (estruturados, semiestruturados e não estruturados) em seu Portal da Transparência. No artigo é detalhado como foi desenvolvido o sistema utilizando a linguagem Java, no entanto não é relatado qual infraestrutura de sustentação para a aplicação. No artigo [20] é proposto um método para projetar modelos de entidade relacionamento (ER) com base no RDF(S) e SPARQL(*SPARQL Protocol and RDF Query Language*). As etapas incluem formulação de regras com a linguagem para consulta de dados SPARQL, análise de recursos e design do modelo ER traduzindo RDF(S), no entanto a discussão fica somente acerca da modelagem, sem apresentar uma infraestrutura proposta. Por fim, o artigo [22] fornece uma explicação de como foi realizado uma interface de pesquisa em RDF utilizando um desenvolvimento em JAVA com a biblioteca Sesane 2 RDF[9] . A biblioteca é um framework de software livre e de código aberto para trabalhar com dados RDF, fornece vários recursos como armazenamento e gerenciamento de dados RDF.A Tabela 1 apresenta os estudos que apresentaram uma solução de infraestrutura para tipo de armazenamento *triplestore*, as tecnologias utilizadas para a infraestrutura e o tipo de ambiente utilizado.

Tabela 1
Relação dos principais artigos e os tipos de infraestrutura utilizada

Artigo	Infraestrutura	Tipo de Ambiente
Zoi Kaoudi [11,12]	*Hadoop*	Cloud
Francesca Bugiotti [13]	EC2 + SIMPLEDB	Cloud
Franke, C [23]	MySQL Cluster + *Hadoop*	Cloud
De Witte, D. [16]	EC2 + S3 + Virtuoso + BlazeGraph	Cloud
Waterloo, G. A. U. [17]	Virtuoso + MonetDB	*On-Premises*
Matono, A. [18]	Postgree + Linux + Apache Jena	*On-Premises*
Victorino, M. [19]	Mysql + Apache Jena	*On-Premises*
Huang, J. [21]	*Hadoop* + RDF+3X	Cloud

[6]https://virtuoso.openlinksw.com/
[7]https://blazegraph.com/
[8]https://geneontology.org/
[9]https://www.w3.org/2001/sw/wiki/Sesame

2.3. Discussão sobre as definições de infraestruturas encontradas na RSL

Esta seção discute as infraestruturas de banco de dados em grafo encontradas na RSL. O *Hadoop* com *MapReduce* [11,12,23,21] oferece escalabilidade e flexibilidade, mas tem implementação complexa e custos elevados na nuvem. O *Amazon SimpleDB* [13] compartilha essas vantagens e desafios. O *HBase* [23] é eficiente na inserção de grandes volumes de dados, mas tem desempenho inferior em consultas e alto custo na transformação de dados RDF. O *MySQL Cluster* [23] oferece bom desempenho em consultas, mas sua escalabilidade é limitada. O *PostgreSQL* [18], embora suporte RDF, tem modelagem complexa e desempenho avaliado em um conjunto de dados pequeno.

3. Contexto de Aplicação: Grafo de Conhecimento para a plataforma ELLAS

A plataforma desenvolvida pela rede de pesquisa ELLAS de pesquisa tem como objetivo providenciar uma maneira de coletar e centralizar dados que auxiliem na compreensão e mitigação da baixa representatividade feminina em cursos e carreiras das áreas STEM (Ciência, Tecnologia, Engenharias e Matemática). Esses dados são espalhados em muitas fontes com diversas granularidades, diferentes línguas, vocabulários e quantidades e na plataforma são conectados para assim, favorecer que pesquisadores e tomadores de decisão produzam estudos para recomendações de políticas para instituições públicas e privadas [24,25]. Tal produção está concentrada em três países da América do Sul (Brasil, Bolívia e Peru) e pretende abordar sistematicamente a questão da liderança feminina em STEM revisando a literatura internacional. A ideia é criar uma plataforma de dados abertos e conectados por ontologias (assegurando sua organização, acessibilidade e reutilização) [27], com o intuito de simplificar e incentivar o uso por qualquer indivíduo interessado no tema. Em resumo, foram definidos dois processos de coleta e estruturação de dados, um para dados não estruturados e outro para estruturados, mais bem descritos em [24, 25, 30, 31]. Os dados não estruturados são organizados em planilhas CSV e passam por etapas de modelagem de ontologias baseada nos dados, validadas em conjunto com os especialistas de domínio do projeto e então são triplificadas e armazenadas no grafo de conhecimento. Já os dados estruturados são coletados de forma automatizada, tendo até o momento como fontes a UNESCO e o INEP (dados estatísticos da educação superior com corte em gênero em cursos de área STEM) [30], e outras fontes serão adicionadas com o tempo.

4. Proposta da Infraestrutura da Arquitetura e Pipeline para a Plataforma ELLAS

As análises realizadas com os resultados da RSL e as principais características do contexto do projeto ELLAS evidenciaram que não há semelhanças com os cenários discutidos nos artigos da RSL e por isso exigiram novas definições para a infraestrutura da arquitetura quanto pipelines para automação da governança de grafos de conhecimento. Dentre as características, destacamos a quantidade de dados que não é tão volumosa, a necessidade de padronizar os dados em sua forma inicial em planilhas CSV devido à equipe de especialistas de domínio possuir familiaridade com planilhas e nenhum conhecimento em grafos de conhecimento nem em ontologias e a necessidade de oferecer visualizações compatíveis com o conhecimento não técnico dos perfis de usuários, que serão tomadores de decisão formadores de políticas para atração de mulheres em carreiras STEM.

4.1. Definições para a Infraestrutura da Arquitetura

A infraestrutura do grafo é gerenciada por um usuário "*Admin*"em um computador local, com armazenamento na AWS. Utiliza-se o EC2 da AWS para o servidor virtual, rodando Windows Server, escolhido pela familiaridade da equipe. O servidor é dedicado ao projeto ELLAS, com um pipeline que transforma

dados CSV em um grafo de conhecimento no *GraphDB* da empresa *Ontotext*[10] . Neste ambiente, está configurada a API para consumo de aplicativos externos e também da plataforma ELLAS. Como já mencionado, nas infraestruturas resultantes da RSL as definições para triplestore se concentravam em grandes volumes de dados, que não é a realidade de muitos projetos medianos de grafos de conhecimento. Buscando por *triplestores* com essas características, foi realizada uma comparação em [28] onde o *triplestore* Graphdb da *empresa* Ontotext apresentou-se como uma solução compatível com o cenário do projeto ELLAS, tornando-se inclusive um case de sucesso reportado no seu site[11] .Na Figura 1, é ilustrada uma versão macro da infraestrutura da arquitetura de governança em que se mostram as definições de ferramental que permite a governança desde o acesso à plataforma pelos administradores dos dados até o usuário final em dispositivos como website e aplicativos.

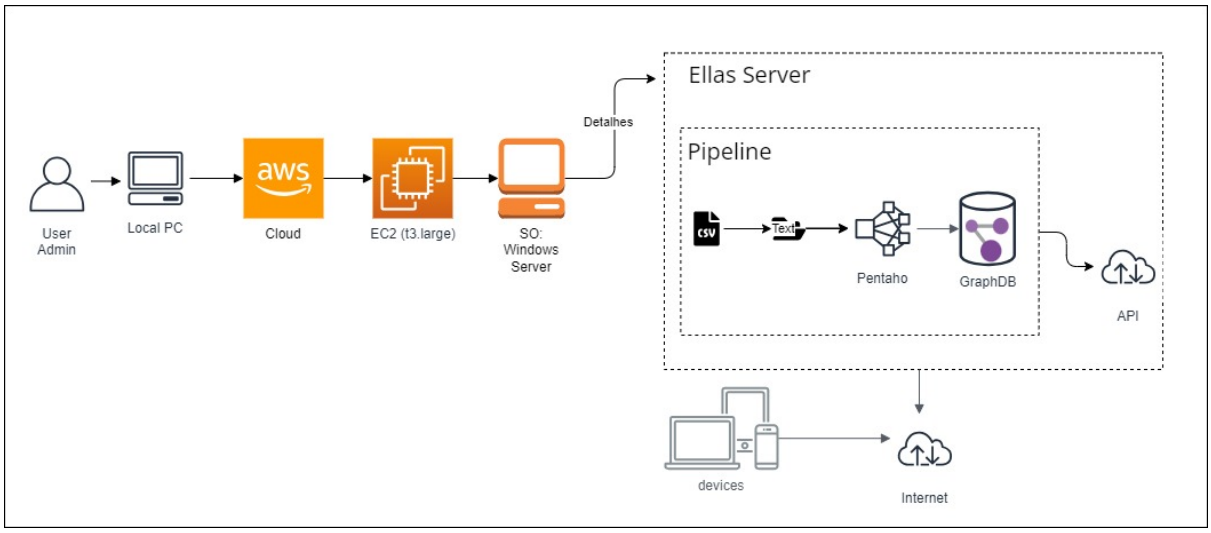

Figura 1: Infraestrutura da Arquitetura de Governança do Grafo de Conhecimento para a plataforma ELLAS. Fonte: o autor.

4.2. Pipeline de Dados

Foi projetado um Pipeline que prevê a manipulação dos dados desde a sua forma crua em formato CSV até a inserção dos dados no triplestore. Para obter um processo de carga de dados de forma automática, foi utilizado o *Pentaho Data Integration*[12] que é uma ferramenta de orquestração de dados de forma visual que combina diversos conjuntos de dados. Na Figura 2 é representado o fluxo de transformação para manipulação dos dados não estruturados (ou dados primários).Os dados brutos em formato CSV são localizados, mapeados e transformados em triplas, para serem inseridos no banco de dados Graphdb em uma rotina de execução pré-estabelecida.

Figura 2: Pipeline automatizado utilizando Pentaho. Fonte: autor

[10]https://www.ontotext.com/products/graphdb/?ref=menu
[11]https://www.ontotext.com/knowledgehub/case-studies/stem-ellas-uses-graphdb-to-help-fight-gender-inequality-in-leadership-positions/
[12]https://pentaho.com/products/pentaho-data-integration/

5. Aplicação e Testes

A aplicação desta infraestrutura, em comparação com os artigos da RSL, é consideravelmente mais simplificada, pois não utiliza camadas de Big Data ou ferramentas mais complexas como o *Hadoop* que são necessárias para a projetos com grandes volumes de dados. A instalação do GraphDB no servidor foi simples, com um arquivo executável com interface conhecida no formato Windows, dispensando configurações avançadas com uso de terminais. Utilizando a licença da Ontotext e seguindo a documentação, foi criada uma API em *Python* para consultas SPARQL. Embora o GraphDB possua um *workbench* para consultas, por segurança e para evitar a exposição do banco, um módulo de consulta em Python foi desenvolvido, controlando as execuções permitidas. Para facilitar a documentação, aplicamos o *Swagger*[13] junto com a biblioteca *Flask*[14] no Python para criar uma página com o serviço de busca para as consultas SPARQL.Como no escopo do projeto temos como premissa uma infraestrutura simples e com um custo-benefício otimizado, na Tabela 2 mostramos o consumo em Dólares (U$) por mês do ano de 2024 desta infraestrutura implementada manipulando um grafo de conhecimento de aproximadamente 50 mil triplas.

Tabela 2
Custo mensal do ano de 2024 da infraestrutura aplicada no Projeto ELLAS.

Mês	Janeiro	Fevereiro	Março	Abril	Maio	Junho	Total
Valor (U$)	11.15	10.91	11.15	8.39	14.23	35.84	91.67

6. Conclusão

A escolha da melhor abordagem para armazenar e gerenciar dados RDF é influenciada por vários fatores, como o volume de dados, o tipo de consultas necessárias, restrições orçamentárias e o nível de familiaridade com a tecnologia e o perfil de pessoas envolvidas na construção e consumo dos grafos de conhecimento. É preciso analisar cuidadosamente as vantagens e desvantagens de cada opção, bem como revisar pesquisas já realizadas para tomar decisões. Foi realizada uma RSL buscando compreender o estado da arte com relação a infraestruturas das arquiteturas desenvolvidas para compreender quais as características dessas arquiteturas bem como dos cenários de aplicação. Para grandes volumes de dados, recursos como Apache *Hadoop* e formatos como *HDFS* foram encontrados, esta abordagem pode ser complexa e onerosa, tornando-se menos viável economicamente. Em um banco de dados de tamanho moderado conclui-se que pode ser utilizado um servidor na Amazon com o serviço EC2, já das opções apresentadas o ideal é trabalhar com um banco de dados de triplestore sem que haja a necessidade de tratamentos em bancos de dados relacionais para armazenar os arquivos RDF e as ontologias. Espera-se que os desafios encontrados pelo time da plataforma ELLAS na definição da infraestrutura da sua arquitetura e pipelines sejam amenizados em projetos semelhantes no futuro para assim fomentar mais e mais o uso de grafos de conhecimento.

Referências

[1] D. Quass, J. Widom, R. Goldman, K. Haas, Q. Luo, J. McHugh, S. Nestorov, A. Rajaraman, H. Rivero, S. Abiteboul, J. Ullman, and J. Wiener. "Lore: a lightweight object repository for semistructured data."*In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 1996.

[2] S. Trißl and U. Leser. "Fast and practical indexing and querying of very large graphs."*In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2007.

[13] https://swagger.io/

[14] https://flask.palletsprojects.com/en/3.0.x/

[3] G. Raschia, M. Theobald, and I. Manolescu. "Proceedings of the first international workshop on open data (WOD)."2012.
[4] Z. Kaoudi, K. Kyzirakos, and M. Koubarakis. "SPARQL Query Optimization on Top of DHTs."*In Proceedings of the International Semantic Web Conference (ISWC)*, 2010.
[5] Z. Kaoudi, I. Miliaraki, and M. Koubarakis. "RDFS Reasoning and Query Answering on Top of DHTs."*In Proceedings of the International Semantic Web Conference (ISWC)*, 2008.
[6] O. Erling and I. Mikhailov. "RDF Support in the Virtuoso DBMS."*Networked Knowledge - Networked Media*, 2009.
[7] S. Harris, N. Lamb, and N. Shadbolt. "4store: The Design and Implementation of a Clustered RDF Store."*In 5th International Workshop on Scalable Semantic Web Knowledge Base Systems*, 2009.
[8] F. N. Afrati, V. R. Borkar, M. J. Carey, N. Polyzotis, and J. D. Ullman. "Map-Reduce Extensions and Recursive Queries."*In Proceedings of the International Conference on Extending Database Technology (EDBT)*, 2011.
[9] Apache Hadoop. "http://hadoop.apache.org/", 2012.
[10] A. Owens, A. Seaborne, N. Gibbins, and M. Schraefel. "Clustered TDB: A Clustered Triple Store for Jena."*Technical Report*, 2008.
[11] Zoi Kaoudi, Ioana Manolescu. "RDF in the Clouds: A Survey."*The VLDB Journal*, 2014. ffhal-01020977ff.
[12] Zoi Kaoudi, Ioana Manolescu. "Cloud-based RDF data management."*In ACM SIGMOD Conference*, June 2014, Snowbird, United States. ff10.1145/2588555.2588891ff.
[13] Francesca Bugiotti, François Goasdoué, Zoi Kaoudi, Ioana Manolescu. "RDF Data Management in the Amazon Cloud."*In Workshop on Data Analytics in the Cloud (DanaC)*, Mar 2012, Berlin, Germany. ffhal-00670492ff.
[14] Melanie Herschel, Ioana Manolescu. "Data Bridges: Data Integration for Digital Cities."*In International Workshop on City Data Management (CDMW)*, November 2012, Maui, HI, United States. ffhal-00757569ff.
[15] E. Meena, A. Kumar, and L. Romary. "An extensible framework for efficient document management using RDF and OWL."*In Proceedings of the Workshop on NLP and XML (NLPXML-2004)*, 2004.
[16] D. De Witte, L. De Vocht, R. Verborgh, K. Knecht, F. Pattyn, H. Constandt, E. Mannens, and R. Van de Walle. "Big Linked Data ETL Benchmark on Cloud Commodity Hardware."*Proceedings of the International Workshop on Semantic Big Data*, 2016.
[17] G. A. Waterloo, G. Aluç, M. Tamer Özsu, and O. M. A. Metrics. "Workload matters: Why RDF Databases Need A New Design."*Proceedings of the VLDB Endowment*, vol. 7, no. 10, 2014.
[18] A. Matono, T. Amagasa, M. Yoshikawa, and S. Uemura. "A path-based relational RDF database."*In Proceedings of the 16th Australasian Database Conference*, 2005.
[19] M. Victorino, M. T. de Holanda, E. Ishikawa, E. C. Oliveira, G. Ghinea, and S. Chhetri. "Proposal of a Brazilian Database Government Open Linked Data: DBgoldbr."*In Proceedings of the 9th International Conference on Management of Digital EcoSystems*, 2017.
[20] L. Xu, S. Lee, and S. Kim. "ER model based RDF data storage in RDB."*In Proceedings of the 3rd International Conference on Computer Science and Information Technology*, 2010.
[21] J. Huang, D. J. Abadi, and K. Ren. "Scalable SPARQL querying of large RDF graphs."*Proceedings of the VLDB Endowment*, vol. 4, no. 11, 2011.
[22] B. Guilfoos, S. Samsi, J. C. Chaves, J. Unpingco, J. Nehrbass, A. Chalker, and A. Krishnamurthy. "Web Interface for Querying/Searching RDF Database."*In Proceedings of the 2007 DoD High Performance Computing Modernization Program Users Group Conference*, 2007.
[23] C. Franke, S. Morin, A. Chebotko, J. Abraham, and P. Brazier. "Efficient processing of semantic web queries in HBase and MySQL cluster."*IT Professional*, vol. 15, no. 3, 2012.
[24] R. C. G. Berardi. "ELLAS Architecture and Process: Collecting and Curating Data on Women's Presence in STEM."*Journal of Information Systems (JIS)*, vol. 15, no. 1, 2024.
[25] C. Maciel, I. R. Guzman, R. C. G. Berardi, N. Rodriguez-Rodriguez, L. Salgado, L. B. Frigo, B. Branisa, and E. Jimenez. "An Open Data Platform to Advance Gender Equality in STEM in Latin America."*Communications of the ACM*, Online First, 2024. doi: 10.1145/3653294.

[26] B. Branisa, P. Cabero, and I. Guzman. "The main factors explaining IT Career Choices of Female Students in Bolivia."*In AMCIS 2021 Proceedings*, 2021.

[27] M. T. de Araújo and A. M. Tonini. "A Participação das Mulheres nas Áreas de STEM (Science, Technology Engineering and Mathematics)."*Revista de Ensino de Engenharia*, vol. 38, no. 3, 2020. Available: http://revista.educacao.ws/revista/index.php/abenge/article/view/1693/905.

[28] G. A. Michelon and R. C. G. Berardi. "Comparative Study of Tools for Modeling, Storage, and Integration of Data on the Semantic Web for the ELLAS Network Platform."*In Proceedings of the WELLAS Conference*, 2023.

[29] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, and A. Zimmermann. "Knowledge Graphs."*ACM Computing Surveys (CSUR)*, vol. 54, no. 4, 2021.

[30] N. Hildebrand, B. Amador, C. Maciel, and R. Berardi. "A Escassez de Dados Abertos Estruturados em Países Latino-Americanos com Enfoque de Gênero na Educação Superior."*In Anais do XVIII Women in Information Technology*, Brasília/DF, 2024, pp. 161-172. doi: https://doi.org/10.5753/wit.2024.2574.

[31] R. Berardi, P. Auceli, C. Maciel, G. Davila, I. Guzman, and 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*, João Pessoa/PB, 2023, pp. 124-135. doi: https://doi.org/10.5753/wit.2023.230764.