Comparative Study of Tools for Modeling, Storage, and Integration of Data on the Semantic Web for the ELLAS Network Platform

Gisane A. Michelon
gisane@unicentro.br
https://orcid.org/0000-0001-6044-3573
Universidade Estadual do Centro-Oeste do Parana, Brazil

Rita C. G. Berardi
ritaberardi@utfpr.edu.br
https://orcid.org/0000-0002-0281-8952
Universidade Tecnologica Federal do Parana, Brazil

Recibido: 21 de agosto del 2023 / Aceptado: 4 de octubre del 2023
doi: https://doi.org/10.26439/interfases2023.n018.6613

ABSTRACT. The layered architecture of the World Wide Web (W3C) defines standards, technologies, languages, and methods needed to build applications that involve the semantic web. From this W3C semantic web architecture was developed the data modeling of the ELLAS (Equality in Leadership for Latin American STEM) network platform. The ELLAS aims to map the policies and context aspects that influence women in STEM (Science, Technology, Engineering and Mathematics) in the countries of Bolivia, Brazil and Peru, which are at different levels of development. The main objective of this work is a comparison of the main storage and semantic integration tools to develop a data model (modeling, storage and integration of data) related to the mapping of policies, initiatives and factors that influence the career development of women in STEM.

KEYWORDS: data integration / semantic web / ELLAS network

ESTUDIO COMPARATIVO DE HERRAMIENTAS DE MODELIZACIÓN, ALMACENAMIENTO E INTEGRACIÓN DE DATOS EN LA WEB SEMÁNTICA PARA LA PLATAFORMA DE RED ELLAS

RESUMEN. La arquitectura en capas de la World Wide Web (W3C) define estándares, tecnologías, lenguajes y métodos necesarios para construir aplicaciones que involucren la web semántica. A partir de esta arquitectura de web semántica del W3C se desarrolló el modelado de datos de la plataforma de la red ELLAS (Equality in

Leadership for Latin American STEM - Igualdad para el Liderazgo en STEM en América Latina). ELLAS pretende mapear las políticas y aspectos de contexto que influyen en las mujeres en STEM (Ciencia, Tecnología, Ingeniería y Matemáticas) en los países de Bolivia, Brasil y Perú, que se encuentran en diferentes niveles de desarrollo. El objetivo principal de este trabajo es una comparación de las principales herramientas de almacenamiento e integración semántica para desarrollar un modelo de datos (modelado, almacenamiento e integración de datos) relacionado con el mapeo de políticas, iniciativas y factores que influyen en el desarrollo profesional de las mujeres en STEM.

PALABRAS CLAVE: integración de datos / web semántica / red ELLAS

ESTUDO COMPARATIVO DE FERRAMENTAS PARA MODELAGEM, ARMAZENAMENTO E INTEGRAÇÃO DE DADOS NA WEB SEMÂNTICA PARA A PLATAFORMA DA REDE ELLAS

RESUMO. A arquitetura em camadas da World Wide Web (W3C) define padrões, tecnologias, linguagens e métodos necessários para construir aplicações que envolvem a web semântica. A partir dessa arquitetura de web semântica da W3C, foi desenvolvida a modelagem de dados da plataforma da rede ELLAS (Equality in Leadership for Latin American STEM - Igualdade no Liderança para Mulheres em STEM na América Latina). O objetivo da ELLAS é mapear as políticas e os aspectos contextuais que influenciam as mulheres em STEM (Ciência, Tecnologia, Engenharia e Matemática) nos países da Bolívia, Brasil e Peru, que estão em diferentes níveis de desenvolvimento. O principal objetivo deste trabalho é realizar uma comparação das principais ferramentas de armazenamento e integração semântica para desenvolver um modelo de dados (modelagem, armazenamento e integração de dados) relacionado ao mapeamento de políticas, iniciativas e fatores que influenciam o desenvolvimento profissional das mulheres em STEM.

PALAVRAS-CHAVE: integração de dados / web semântica / rede ELLAS

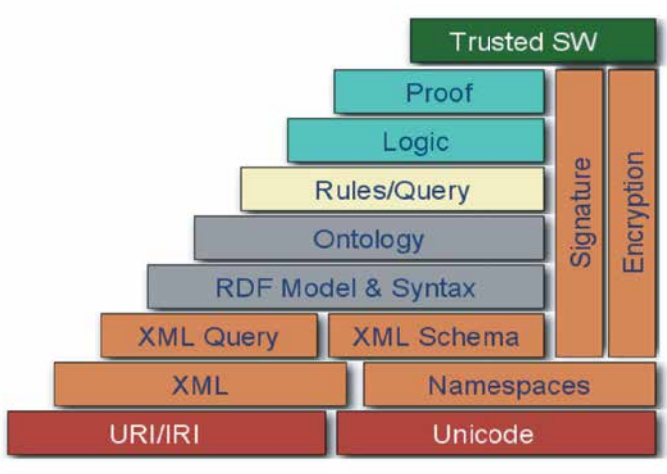
1. INTRODUCTION

The open and connected data platform of the ELLAS[1] research network aims to map existing policies and initiatives in the context of female presence in STEM (Science, Technology, Engineering, and Mathematics) courses and factors influencing women in STEM in Bolivia, Brazil, and Peru, countries at different levels of development (IDRC, 2023). Various factors can influence women in their choice of field of study, one of which could be the experience in school and the stereotyped discourses of their teachers (Berardi et al., 2022). An open data platform, initially available in three languages (Spanish, Portuguese, and English), will be built to map these informations and enhance collaboration between the education sector, government, and industry in efforts to reduce the STEM gender gap in Latin America, increasing the number of female leaders in universities, industries, and public institutions. The platform will use semantic web technologies to structure data in an integrated manner through ontologies.

The World Wide Web Consortium (W3C) proposed a layered architecture for building applications that involve the semantic web. This architecture defines the standards, technologies, languages, and methods necessary to make web resources available including for machines (Cantele, 2009; Cantele, 2017). Figure 1 presents this architecture, proposed by the composition of layers, here highlighted between i) ontology and inference layer and ii) data layer.

Figure 1

Architecture of the Semantic Web



Note. Adapted from W3C (2004)

1 https://ellas.ufmt.br/

The data layer defines both structured and unstructured data in various formats (Comma Separated Value - CSV, Structured Query Language - SQL, etc.) that need to be mapped to the RDF (Resource Description Framework) language, which structures data into triples. For this purpose, tools such as OpenRefine[2], OntoRefine[3] e karma[4].

The ontology and inference layer involves the definition of ontology, queries in the SPARQL language, inference, and security. Tools identified in the literature as potential options for use in this layer include:

- For Ontology Definition: OntoWiki[5], SMW+[6], Neon[7], Protégé[8], Webprotégé[8] e Swoogle[9].
- For the storage of triples: Virtuoso (DB-Engines, 2022), GraphDB (Ontotext GraphDB, 2022), AlegroGraph[10], Apache Jena Fuseki[10] e Neo4J[10].
- Semantic Integration (para consultas federadas): Semagrow (Semagrow, 2023), DARQ (DARQ, 2023), SPLENDID (Görlitz and Staab, 2011) e FedX (Ontotext GraphDB, 2022).

As there are numerous tools that can be used in each layer and there is no consensus on which tools are most suitable according to the application domain, a study and comparison of the main tools of the data storage and semantic integration layers were necessary to determine which would be used in the data modeling stage of the ELLAS research network platform[11].

In this context, the goal of this work is to conduct a comparative analysis of tools to enable the definition of a framework of tools that meet the needs and specificities of the platform to be built. These needs range from ontology modeling, data integration, to web publication. For this purpose, a comparison of the tools was carried out through articles that theoretically address the characteristics of the tools, as well as articles that have used these tools in their applications.

The rest of this text is organized as follows: Section 2 discusses the development of the work, Section 3 addresses the results achieved and discussions, and Section 4 describes the conclusions.

2 https://openrefine.org/
3 https://www.ontotext.com/products/ontotext-refine/
4 https://usc-isi-i2.github.io/karma/
5 http://docs.ontowiki.net/
6 https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki_Plus
7 http://neon-toolkit.org/
8 https://protege.stanford.edu/
9 https://dept.abcdef.wiki/wiki/Swoogle
10 https://db-engines.com
11 https://ellas.ufmt.br/

2. DEVELOPMENT

Through bibliographic research, in scientific articles and initiatives that collect and provide information on database management systems (DBMS), such as DB-Engines (Knowledge Base of Relational and NoSQL Database Management Systems) (DB-Engines, 2022) and Capterra (Capterra Inc, 2023), it was noted that two of the most used tools in graph databases are Virtuoso and GraphDB.

For the comparison of triplestore database management systems, the demands that the ELLAS network platform required for the project were taken into consideration. Due to these demands, the criteria chosen for the comparison are:

- Model of an exclusive graph database;
- Storage of triples and also support for quadruples;
- Open source license or a free version;
- Variety of programming languages, operating systems, and APIs supported;
- Intuitive interface;
- Supported sizes of triples.

In semantic integration/federation, the tools compared are Semagrow, FedX, SPLENDID, and DARQ, as they are among the most cited tools in the literature of the field. Bibliographic research was conducted in scientific articles, such as: (Charalambidis et al., 2015), (Saleem et al., 2018), (Charalambidi et al., 2015), (Rolim et al., 2021). The web pages of the tools available online were also researched (Semagrow, 2023), (Ontotext GraphDB, 2022), (DARQ, 2023).

The work of Rakhmawati and Hausenblas (2012) served as a basis for comparison, in relation to the criteria used. However, the federators addressed were not the same as those in this work, such as Semagrow. Therefore, it was decided to create a comparative table that covered all the federators, objects of this work (DARQ, Semagrow, FedX, and SPLENDID), along with other criteria based on (Charalambidis et al., 2015), (Saleem et al., 2018), (Charalambidi et al., 2015), (Rolim, 2021), (Semagrow, 2023), (Ontotext GraphDB, 2022), (DARQ, 2023).

3. RESULTS AND DISCUSSIONS

Table 1 presents a comparison between the storage tools, GrahpDB and Virtuoso, which are currently the most used tools in practical projects. However, even though these tools are the most cited, no explicit indication was found in the literature regarding the advantages of one over the other in practical projects. In this sense, the two storage tools were compared based on the database model, supported operating systems, access methods, supported programming languages, inference models, ease of working with the interface, and the amount of supported triples.

Table 1

Comparison of GraphDB and Virtuoso

Characteristics	GraphDB	Virtuoso
Primary DB Model	Graph DBMS RDF store	Document store Graph DBMS Native XML DBMS Relational DBMS RDF store Search engine
License	Commercial*	Open source
Supported Operating Systems	All with Java VM Linux OS X Windows	AIX FreeBSD HP-UX Linux OS X Solaris Windows
APIs and Other Access Methods	GeoSPARQL GraphQL Federation Java API JDBC RDF4J API RDFS RIO Sail API Sesame REST HTTP Protocol SPARQL 1.1	ADO.NET GeoSPARQL HTTP API JDBC Jena RDF API ODBC OLE DB RDF4J API RESTful HTTP API Sesame REST HTTP Protocol SOAP webservices SPARQL 1.1 WebDAV XPath XQuery XSLT
Supported Programming Languages	.Net C# Clojure Java JavaScript (Node.js) PHP Python Ruby Scala	.Net C C# C++ Java JavaScript Perl PHP Python Ruby Visual Basic
Supported Size	10 Billion triples	50 Billion triples
Ease of Working with Interface	Yes	No

Note. Adapted from DB-Engines (2022)

According to research in articles that conducted tests, Virtuoso performs better than GraphDB, but has a poor interface (Rosa et al., 2019), (Addlesee, 2018). Since the data volume for the ELLAS network platform is not very significant, performance is not a determining factor. According to the comparison table, both tools are quite efficient and similar to

each other, with similar supported programming languages, as well as operating systems. However, the use of Virtuoso in a Windows environment does not seem to offer ease of use in its interface, which makes GraphDB appear as a better option. Additionally, the amount of triples supported by the free version of GraphDB is sufficient for the platform. Therefore, for the ELLAS network platform, GraphDB was chosen for its more intuitive interface, the team's existing experience with it, and its ease of use in federation.

For a single ontology to have direct access to multiple data sources, federation can be used. Federators allow the execution of SPARQL queries across multiple local knowledge graphs (federated query) (Rolim et al., 2015). In the development of the ELLAS network platform, several knowledge graphs will be used to integrate data from different sources, which can be integrated into a single ontology..

In Table 2, a comparison is made between the most commonly found semantic integration tools in the literature: Semagrow (Charalambidis et al., 2015), FedX (Rakhmawati and Hausenblas, 2012), SPLENDID (Görlitz and Staab, 2011) e DARQ (Saleem et al., 2018). The comparison is based on the criteria of source selection strategy, type of information collection, culminating in advantages and disadvantages.

Table 2

Comparison of Federators

Federator/ Feature	FedX	Semagrow	SPLENDID	DARQ
Data Source Selection Strategy	Non-catalog-based: Uses an on-the-fly technique where source selection is based on ASK queries. Initially, the data source list lacks statistical information. ASK query is cached.	Data catalog: Uses VOID metadata to optimize queries generated through statistics obtained directly from the data.	Data catalog: Utilizes statistics from VOID vocabulary descriptions (at system load). ASK query is sent to each dataset for verification.	Data catalog: Uses service description (data and statistical information) to decide where to send a subquery. With the service's predicate list, it plans the query. Catalogs integrate subquery results.
Type of Information Collection	Real-time and cache	-	Real-time and cache	Cache
Advantages	Efficient query execution techniques (semi-joins). Does not use statistics for query optimization, relies on join order heuristics. All patterns used are supported by current data sources.	Query optimizer introduces little overhead and works in the absence of metadata (has appropriate fallbacks).	Exclusively depends on VOID statistics, thus can integrate almost any RDF data source.	-

(continúa)

(continuación)

Federator/ Feature	FedX	Semagrow	SPLENDID	DARQ
Disadvantages	Accesses all data only through GraphDB. DESCRIBE queries are not supported. FedX is not stable with queries like {?s?p?o} UNION{?s?p1?o} FILTER(xxx). The federation only works with remote repositories.	Difficult configuration. Only works on Linux. Requires Apache Tomcat. No recent updates.	Uses catalogs that directly decide how to distribute subqueries (it can be assumed there's a difference between the graph and class partitions).	Requires proprietary extensions to protocols not supported by most current endpoints. Like SPLENDID, by using catalogs, it may decide how to distribute subqueries.

The choice of the tool to use in the federation was FedX, as it collects information in real-time and cache, is faster than SPLENDID (Charalambidis et al., 2015), and does not require proprietary extensions for unsupported protocols on endpoints. Also, because Semagrow requires configuration effort and recent updates are not being provided. Additionally, FedX allows for easier extension of the model to federate with GraphDB, as both tools are from Ontotext.

An alternative to using federators is internal federation. This type of federation assumes the existence of a repository for each graph. It uses the SERVICE proposal but, in this case, executed on the same SPARQL Endpoint with distinct repositories.

GraphDB utilizes the concept of repositories, where each repository can have 1 or N graphs stored as 'named graphs'. This allows for the possibility of having one graph per data source, with the option of having all graphs with data from sources in the same Endpoint. This alternative is better in terms of query cost since it operates on the same SPARQL endpoint and repository, as described in (Ontotext GraphDB, 2021) and translated: "The HTTP transport layer is bypassed, and iterators are accessed directly. The speed is comparable to accessing data in the same repository."

If, in the future, it becomes necessary to use a single ontology with project updates, you can opt for federation, which is why FedX was chosen.

4. CONCLUSION

The main objective of this work was the study and comparison of data storage and integration tools aimed at establishing a data model related to mapping factors that influence the career development of women in STEM, in the early stages of the ELLAS network platform. It is believed that with this comparison, other projects can benefit to assist in deciding which tool is most suitable for their context. For the context of the platform developed by the ELLAS network, the GraphDB and FedX tools appear to be more suitable. As future work, we intend to use the tools and evaluate if they were indeed suitable and contribute lessons learned about their use.

REFERENCES

Addlesee, A. (2018). *Comparing Linked Data Triplestores*. Medium. https://medium.com/

Berardi, R. C. G., Amador, B. O., Hoger, M. D. V., Turato, P. A., Santos, L. M. da S., & Bim, S. A. (2022). The demand for stereotype-free computing courses for Elementary School Teachers. *Journal on Interactive Systems*, *13*(1), 410-418. https://doi.org/10.5753/jis.2022.2854

Cantele, R. C. (2009) *Construindo Ontologias a Partir de Recursos Existentes: Uma Prova de Conceito no Domínio da Educação* [Tese de doutorado, Escola Politécnica da Universidade de São Paulo]. Departamento de Engenharia de Computação e Sistemas Digitais, São Paulo.

Cantele, R. C. (2017). *Machine Learning: sistemas baseados em regras*. iMasters. https://imasters.com.br/

Capterra Inc. (2023). *Discover the resources you need to grow your business*. https://www.capterra.com

Charalambidis, A., Troumpoukis, A., & Konstantopoulos, S. (2015). SemaGrow: Optimizing federated SPARQL queries. In *Proceedings of the 11th International Conference on Semantic Systems* (pp. 121-128).

Charalambidis, A., Konstantopoulos, S., & Karkaletsis, V. (2015). Dataset descriptions for optimizing federated querying. *Proceedings of the 24th International Conference on World Wide Web.*

DARQ. (2023). *DARQ - Federated Queries with SPARQL*. https://darq.sourceforge.net/

DB-Engines. (2022). *System Properties Comparison GraphDB vs. Virtuoso. DB-Engines*. https://db-engines.com/en/

Görlitz, O., & Staab, S. (2011). SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. In *Proceedings of the Second International Conference on Consuming Linked Data* - Volume 782, 13-24.

IDRC. (2023). *Latin American open data for gender-equality policies focusing on leadership in STEM*. IDRC CRDI. https://idrc-crdi.ca/en

Ontotext GraphDB. (2021). *Internal SPARQL Federation*. GraphDB. https://graphdb.ontotext.com/documentation/9.6/standard/index.html

Ontotext GraphDB. (2022). *FedX Federation*. GraphDB. https://graphdb.ontotext.com/documentation/10.0/fedx-federation.html

Rakhmawati, N. A., & Hausenblas, M. (2012). On the impact of data distribution in federated SPARQL queries. In *IEEE Sixth International Conference on Semantic Computing, Palermo, Italy* (pp. 255-260). https://doi.org/10.1109/ICSC.2012.72

Rolim, T., Avila, C., Mariano, R., Calixto, T., Ivo, P., Filho, J., Brayner, A., & Vidal, V. (2021). Uso das Tecnologias da Web Semântica na Construção de Grafos de Conhecimento Semântico baseado no Enfoque Híbrido. En *14º Seminário de Pesquisa em Ontologias no Brasil – ONTOBRAS*.

Rosa, F. L., da Silva Machado, R., Primo, T. T., Yamin, A. C., & Pernas, A. M. (2019). Análise de desempenho de ferramentas para persistência de dados ontológicos em triplas: Experimentos e resultados. In *XII Seminar on Ontology Research in Brazil* (pp. 1506-1509).

Saleem, M., Khan, Y., Hasnain, A., Ermilov, I., & Ngonga Ngomo, A. C. (2018). An evaluation of SPARQL federation engines over multiple endpoints. In *International Semantic Web Conference*. NUI Galway.

Semagrow. (2023). *Página oficial do Semagrow*. http://semagrow.github.io

W3C. (2004). *Developing Core Web Services Standards at the W3C*. https://www.w3.org/