Daniela Cristina Bula Gonçalves Balaniuc

Semantic Web

São Paulo

2023

Daniela Cristina Bula Gonçalves Balaniuc

Semantic Web

Course completion work for
obtaining the degree in
Specials in Data Science
presented to the Anhembi
University
Morumbi.
Advisor: Professor Rafael Guem
Murakami

São Paulo
2023

Daniela Cristina Bula Gonçalves Balaniuc

Semantic Web

Course completion work for

obtaining the degree in

Specialization in Data Science

presented to the Anhembi University

Morumbi

To my relatives,
with affection.

THANKS

To God, for being always in my life and for having given strength so that I could complete the course.

To my family, for encouraging me to continue my studies. Professor Rafael Guem Murakami, for the guidance given in the preparation of this work.

"Persistence is the path to success." (Chales Chaplin)

ABSTRACT

The case study aims to solve the problem of chronic patients with blood glucose in the health system of the city of Winderson with approximately 15,000 inhabitants standing out with a population over 40 years, management carried out through spreadsheets of data time to return information about patients, impairing the understanding by health professionals, to improve the information of patients by organizing the information in a readable way, and performing the best form of the concepts. This paper proposes a case study that will serve as a basis for a framework that supports the evolution of the semantic web in the health field. The Semantic Web is a tool used in data manipulation enabling the search for information in an organized way to facilitate the prevention and treatment of patients with diabetes. Aiming to provide a basis of the concepts of the Semantic Web, in 5 stages with examples of development environments and implementation of applications for the manipulation of data published on the Web. The stages of development are presented in: 1. Understanding the Semantic Web 2. Semantic Web and Standards 3. The Semantic Web Application. 4. Use of Data Warehouse. 5. Use of RDF (Resource Description), Graphs and SPARQL (Protocol and RDF Query Language).

Keywords: Semantic Web. Semantic Web and Standards. The Semantic Web Application. Use of Data Warehouse. RDF (Resource Description). Graphs and SPARQL (Protocol and RDF Query Language).

List of Illustrations

Figure 1 - Semantic web architecture…….….14

Figure 2 - W3C Chet Sheet..... 16

Figure 3 - Semantic of Information Connections..... 17

Figure 4 - User Interface & Applications..... 17

Figure 5 - Ontology application and stored as RDF triples (triple warehouse) 18

Figure 6 - Semantic Database..... 19

Figure 7 - Resource Descriptor Framework (RDF)……..... 20

Figure 8 - Web of Documents to Web of Data……..... 20

Figure 9 - framework to be used in cases of diabetes by age and localization.....22

Lista de Tables

Tabela 1 - Resource Descriptor Framework (RDF)..... 20

LIST OF ABBREVIATIONS AND ACRONYMS

APIs - Interface of the Application Program

DW – Data Warehouse

MD - Multidimensional

HTML – Hypertext Markup Language

OLAP – Online Analytical Processing

OWL – Language of Web Ontology

RDF – Resource Description

RDFS - Taxonomies

RIF/SWRL - Rules

SQL – Structure Query Language

UNICODE – Set of Characters

URI – Uniform Identifier of Resources

XML – Extensible Markup Language

SUMMARY

1 INTRODUCTION 13
2 DEVELOPMENT 13
3 FIELD RESEARCH 23
4 CONSIDERATIONSFinais..... 24
IMPROVEMENTS 25

INTRODUCING

With the health system of the city of Winderson with approximately 15,000 inhabitants standing out with a population over 40 years and with many chronic patients with blood glucose. Relying on the semantic web, allowing indexing meaning to words so that they can be understood by machines and humans in order to meet people and not computers, data in a format that allows computer systems to understand the content. The internet provides a freedom to its users, in the same environment having sophisticated websites, built with the help of experts, whose training serves as a reference for companies throughout the country, cohabiting with personal pages, may be a facilitator in the area of health. Highlighting the existence of some layers of technologies and web standards implemented in the architecture present in the semantic web. Using information search in an organized way by the semantic web, the semantic web architecture formed from the layers Unicode and URI (Uniform Resource Identifier), XML (Extensible Markup Language), RDF (Resource Description Framework) Schema, ontology, digital signature, logic, proof and validation is essential for indexing, words and information handling identifying patterns as languages and methodologies in the RDF (Resource Description Framework) and the OWL (Ontology Web Language)The application of the Semantic Web is used to visualize the applicability in the health area.

DEVELOPMENT

1. Semantic Web

First, building a connected database, Linked Data, presented by Tim Bernes-Lee, the World Wide Web (W3C) creates an efficient way to represent data, such as a web of information built on a global scale to be processed by machines. The central idea is to categorize the information in a standardized way, facilitating its access. This idea is similar to the solution used for the classification of living beings, for example the well-defined taxonomy, adopted and shared by most researchers in the world. According to Tim Berners Lee the Web

Semantics should be as decentralized as possible. Because it has an architecture present in the semantic web, it is essential to identify and understand what each layer can rally to make a correct and sensible use, so the following are the layers; The Unicode layer refers to a set of objects that define names and resources in the semantic web. The XML layer (Extensible Markup Language) corresponds to a computational language

for structuring data, the Namespace layer corresponds to a collection of names used in XML documents (Extensible Markup Language), which are intended to validate elements and attributes. The Resource Description Framework (RDF) Core layer provides the technology for semantic modeling, which can create computational linkages. The RDF (Resource Description Framework) Schema corresponds to a semantic extension of the RDF (Resource Description Framework) code. DLP is a technology that performs the intersection between paradigms that represent knowledge, which are descriptive logic and logical programming. In the signature layer corresponds to the set of technologies that perform the signature function of a determined person. In encryption there is a process of encoding or embedding information, which is intended not to be interpreted by unauthorized persons. The SPARQL layer (Protocol and RDF Query Language) is a computational language similar to SQL, intended for queries. And the OWL layer (Web Ontology Language) corresponds to a computational language for ontology development. The rules allow the definition of information resources. In the logical framework layer allows the definition of rules to be treated the information that was described in the previous levels. The proof layer makes it possible to check for consistency or logical proof of resources. And the validation layer, related to confinement and designed to ensure that information is represented correctly. The following figure shows the architecture of the semantic web.

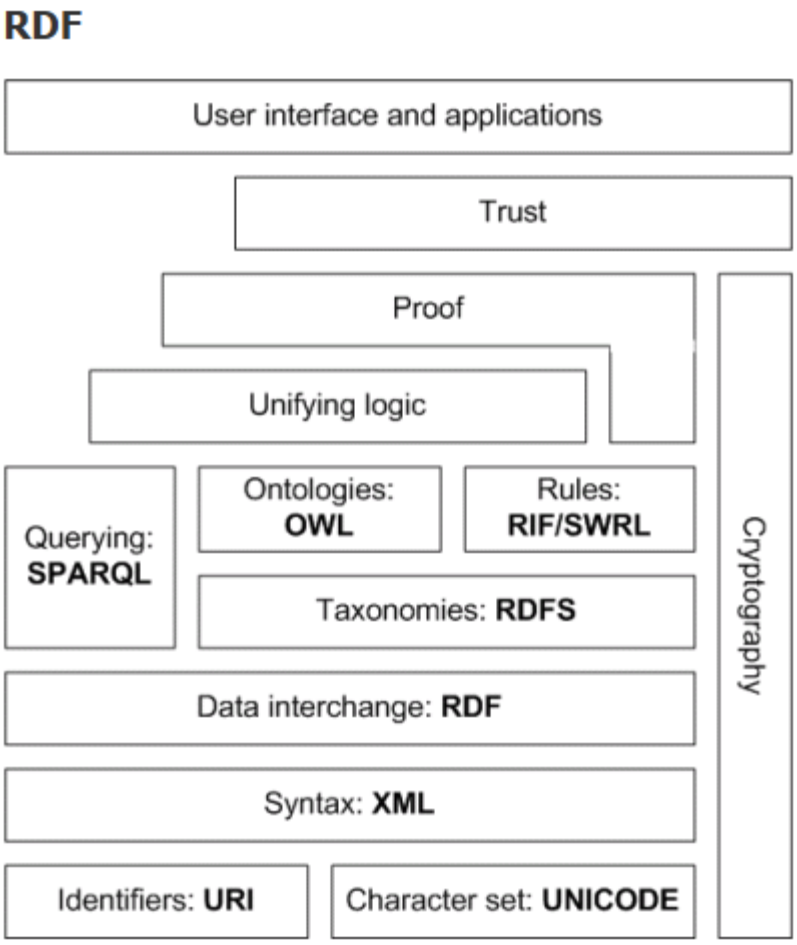

Figure 1 - Semantic web architecture

2. Web Semantics and Standards

Second, the semantic web and its patterns, is known as "web of documents", aiming to enable searches in a database, enables computers to have a system to support interactions on the network, able to create repositories of data on the Web, construct vocabularies and write rules to interoperate with this data.
In the semantic web it is possible to have data connected. This is Linked Data, referring to a set of best practices for publishing and connecting structured data on the Web, allowing to establish links between items from different data sources to form a single space

data, is the web of data connecting the links referring to dates, titles; numbers, chemical properties and any other data that can be conceived. The Resource Description Framework (RDF) is the basis for publishing and linking data. Other technologies allow you to insert data into documents (Annotation website with RDF data), GRDDL (Aware agent, extract all hard data from pages that Reference that link)) or expose what you have in a SQL database. Vocabularies and ontologies are important tools for organizing domain data. Queries using the SPARQL (Protocol and RDF Query Language). Inference for reasoning about the data from

the rules. Vertical applications for example the W3C work in different sectors such as health, government and energy. In the area of Health, the semantic web assists decision-making in the field of clinical research, and to interoperate medical information between institutions. Patterns are used in many areas as 'models' or abstract descriptions that bring together the best practices of some field, in computing is usually attributed to inspiration drawn from the field of architecture, design Patterns as software standards. Os padrões trata a complexidade constituída pelo entrelaçmento das questões culturais, sociais e históricas que vão além dos estudos desenvolvidos e apontam para um consequente sistema de alta complexidade tornando um problema com diferentes facetas e constituindo uma relação em uma estrutura de ligação que consiga apresentar e relacionar os dados.

Os modos de visualização e recomendações do W3C segue como: por tecnologia (como “todas as especificações relacionadas a HTML (Hyper Text Markup Language)”), por status (ordenada inicialmente pelas mais recentes e, em seguida, agrupadas por status, das que já são Recomendação até as que ainda estão no status de Rascunho), por data(a partir das mais recentes), por título, por grupo (os grupos que trabalham nas especificações)

Aqui temos um link aonde é possível fazer buscasnos padrões da web semântica: search

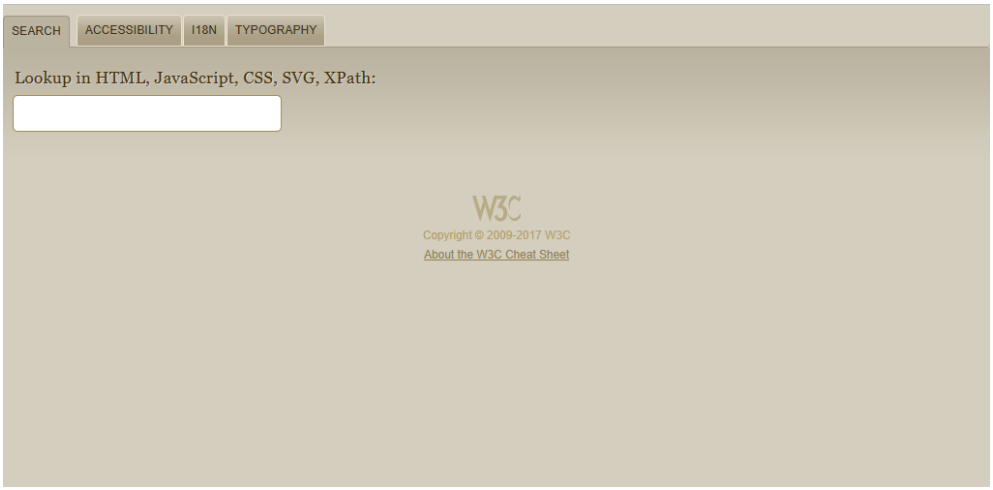

Figura 2 - W3C Chet Sheet

With the evolution of the connections of how the web was made starting with the first protocols used only for email, specific files like PC Era,

Web1, Web2, Web3, Web4. See in figure 3 the Timeline map.

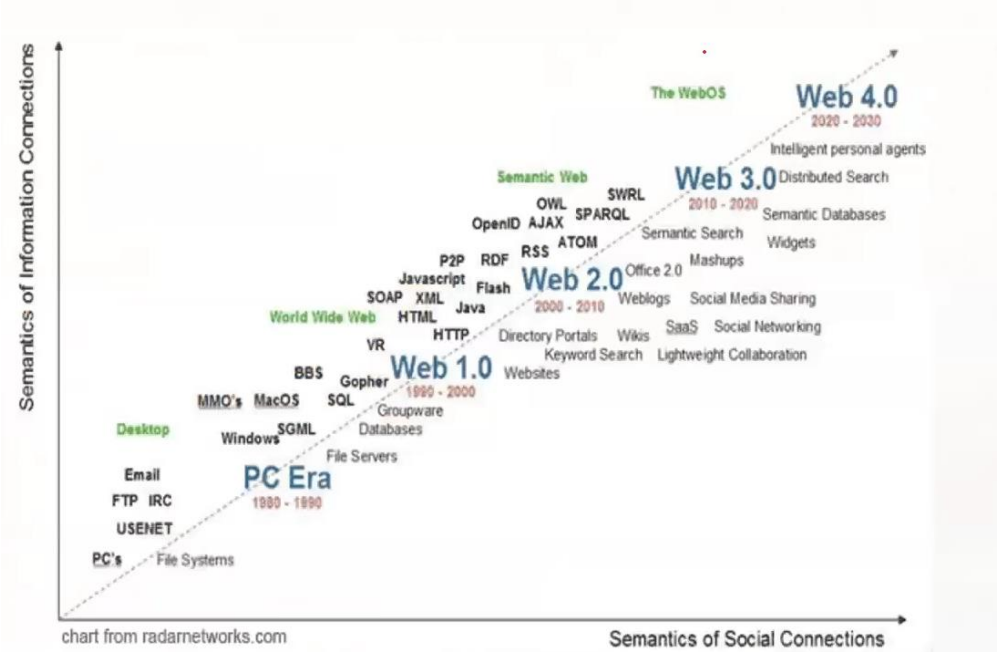

Figura 3 – Semantic of Information Connections

3. The Application of the Semantic Web

Third, for the concept of the World Wide Web based on public blockchains, the semantic web also known as Web3 from the World Wide Web interaction, uses web design and application referring to the standards of web page development. In linked data applications, there are a few categories: Linked data manager, linked data consumer, and web user interface. The linked data consumer retrieves linked data from data sources, if the recovered data is not in RDF format, Wrappers can be used to translate the data into linked data. Systems that consume only linked data are usually called mashups. The Linked Data consumer is responsible for manipulating the Linked Data consumed to produce new Linked Data. The user interface provides a way to interact with the application. With three main types of ratings: 1. Linked data browsers consume linked data and present in a way that allows users to browse it. 2. Linked data search engines. 3. Domain-specific linked data applications. Applications of Linked Data may use Semantic Web technologies either extrinsically or intrinsically. If semantic web technologies are used extrinsically, linked data is consumed and processed using APIs. Traditional technologies such as Relational Database Management Systems (RDBMS) can be used for internal storage processing. In the intrinsic use of Semantic Web technologies, it can be stored in the internal state of the application in a

triplestore instead of using RDBMS.

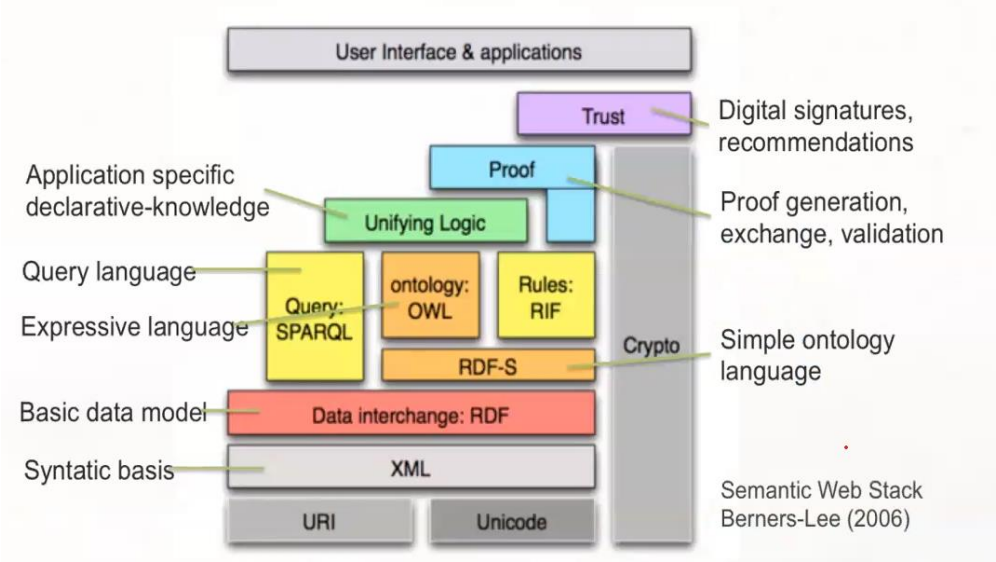

Figura 4 – User Interface & Applications

4. Use of Data Warehouse

However, data Warehouse is a decision support tool that collects your data from operational databases and various external sources, transforms information into information and makes that information available to decision makers in a consolidated and consistent manner. OLAP (Online Analytical Processing) is a technology associated with DW (Data Warehouse) processing in multidimensional (MD). Analyze and explore large amounts of semantic data by combining the inference power of annotation semantics with analytics capabilities provided by aggregations, navigation, and OLAP-style reports. Semantic data must be organized in a well-defined conceptual MD (multidimensional schema, so that sophisticated queries can be expressed and evaluated. In this architecture, the integration of the data model is defined as an ontological application that models the health scenario (patients, visits, reports, etc.) Figure 5 - Biomedical data is semantically annotated according to the ontology application and stored as RDF triples (triple storage).

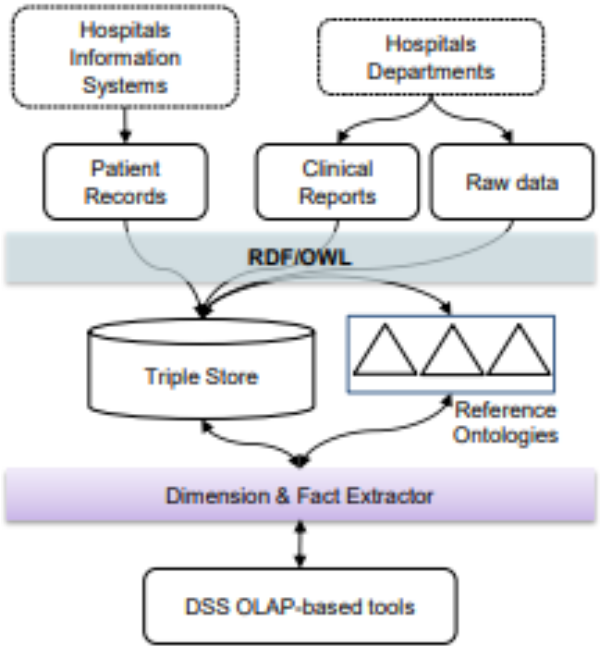

Figure 5 - Ontology application and stored as RDF triples (triple warehouse)

5. Use of RDF (Resource Description), Grafos e SPARQL (Protocol and RDF Query Language).

Then the RDF (Resource Description) works through the triple relation between resource, being composed of a subject, the predicate of a statement (propriety) and the value, which is the object of a statement. Offering subsidies to build semantics from the interconnected data and is the main technology that underlies the semantic representation of web. Using XML (Extensible Markup Language) and schema validations enabling the correct use of markings. The RDF (Resource Description) has several characteristics, such as the description of declarations through a mechanism called rectification, which allows a statement to have treated co-ownerships such as: a subject, which defines the resource being described by the modeled statement; the predicate, which identifies the original property of the modeled statement; the object, which identifies the property value in a modeled statement; the type, which describes the type of the new resource. The SPARQL (Protocol and RDF Query Language) and Graphs are technologies that assist in the visualization of the results. First, I will identify graphs, which are widely used in both mathematics and programming, defining a graph as a set of vertices, also known as nodes, which are interconnected by lines, known as edges. In the semantic web the connection of graphs is ugly with concepts of Uniform Resource Identifier (URI). From this link, so the vertex (node) being unique, as having a URI with a Wikipedia Diabetes disease - Wikipedia and whenever you want to cite diabetes, uses this Uniform Resource Identifier (URI) as the main point. If we connect all the databases with semantic information about the word diabetes, we will have a graph similar to the following image, or is, a graph with several vertices and edges, with

various information. See figure 6 - Semantic database.

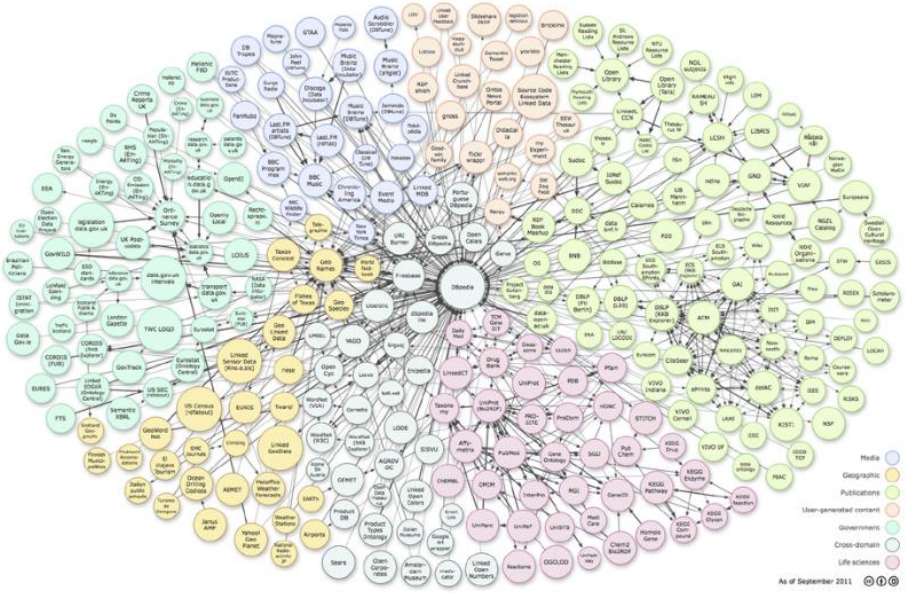

figure 6 - Semantic database

As a result of RDF this is an example in the search of the Resource Descriptor Framework (RDF) in Figure 7 - Resource Descriptor Framework (RDF).

```
<rdf:RDF>
  <viaf:NameAuthorityCluster rdf:about="viaf/12345679> …
  <viaf:EstablishedHeading
    rdf:about="viaf/12345679/#Mozziconacci,+Jean-François“>
  <viaf:NameAuthority
    rdf:about="viaf/12345679/#LC%7Cn++93057547“>…
  <foaf:Person rdf:about="viaf/12345679/#foaf:Person“>…
  <skos:Concept rdf:about="viaf/12345679/#skos:Concept“>…
</rdf:RDF>
```

Figure 7 - Resource Descriptor Framework (RDF).

Currently in Web 3.0 (semantic), data from Documents of diseases like diabetes and a mixture of information from patient data in Hyperlinks, following to the web of data with information from many data Typed Links, as well as a person can combine information from many text documents to solve a problem or perform a task. Also being able to see future stages in the development of the semantic Web, including an additional stage defined Web 4.0. See Figure 8 - Web of Documents to Web of Data.

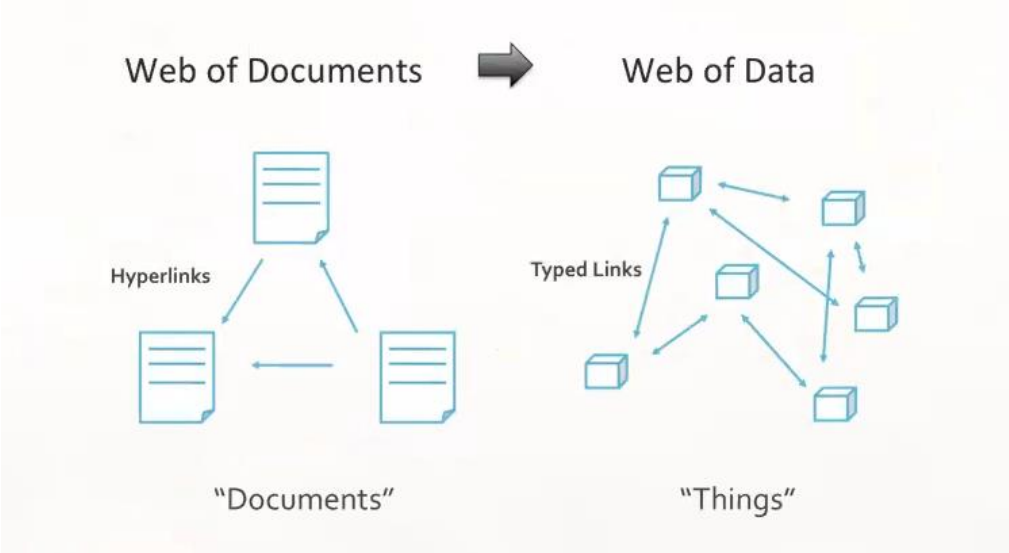

Figure 8 – Web of Documents to Web of Data.

Consequently, the SPARQL (Protocol and RDF Query Language) knowing that in the semantic web, the data is a conceptual data model of RDF, which can be stored in a relational database with RDF mapping schema and others. This language can be compared to the linkage of databases already learned by many, the SQL query language (Structure Query Language). Similar to SQL (Structure Query Language), it can handle results from SPARQL queries such as: LIMIT: limits the number of rows; DISTINCT: removes duplicate rows; ORDER: sorts the result. Having an importance in the application of the techniques of the semantic web, it allows query in data published in RDF and Web Ontology Language (OWL). The Ontology is aggraded and facilitates the realization of research in the most diverse materials already existing, and can help in the cure of certain diseases such as the study of diabetes.

Finally, the most commonly used multitier architecture in three layers. A presentation layer provides a user interface that can accept user input and render the results in a human-readable format. The logical layer implements the business logic of the application. This takes the data and analytics available and transforms them to meet user needs. The data layer stores the underlying data in a form independent of the business logic applied to it in the application. Figure 9 illustrates the example of diabetes with a three-layer architecture. The presentation layer handles user queries and returned outputs, including textual results and views. The logical layer transforms user queries into SPARQL queries and aggregates RDF results. The data layer is responsible for storage and in this case uses a triple store. An important aspect to note in the case of linked data applications is that the dividing line between the data layer and the logical layer may not be as clear. If a relational database is used as a storage tier, all data processing in addition to the return of results to a database query is done at the logical tier. However, triple stores are capable of performing various types of reasoning and therefore in some cases a significant part of business logic can be realized within the triple store. For performance reasons, it is

advantageous to perform reasoning at the lowest possible level. See figure 10 - framework to be used in cases of diabetes by age and location.

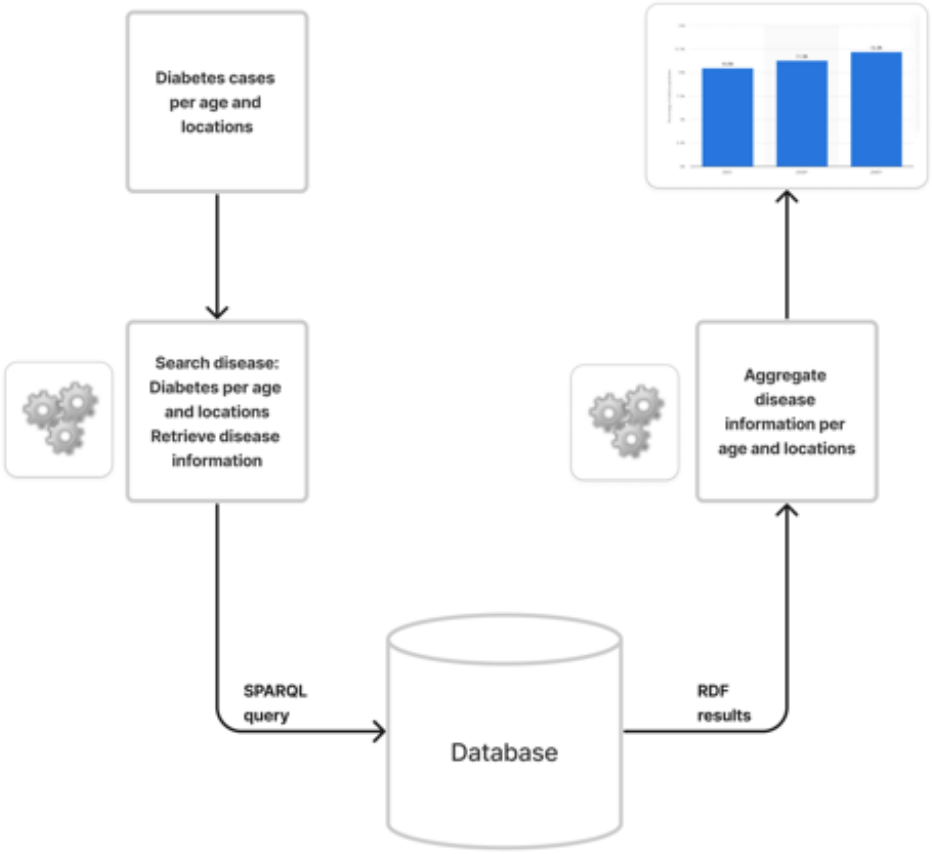

Figure 9 - framework to be used in cases of diabetes by age and location.

CONCLUSION

Concluding, the case of the health of chronic patients with blood glucose in the health system of the city of Winderson with approximately 15,000 inhabitants with a population above 40 years. More and more semantic data is becoming available on the web with various initiatives that promote a change in the current Web towards the Web of Data, where data semantics becomes explicit through data representation formats and standards such as RDF/(S) and OWL. However, this initiative had not been accompanied by efficient intelligent applications being able to exploit implicit semantics and thus provide more insightful analysis. This work proposes the basis of the construction of the framework created and supporting the evolution of the semantic web in the health domain. The Semantic Web, a tool used in data manipulation enabling the search for information in an organized way to facilitate the prevention and treatment of patients with diabetes. Aiming to provide a basis of the concepts of the Semantic Web, in 5 steps with examples of development environments and implementation of applications for the manipulation of data published on the Web. The development stages of the Framework were developed using the following steps: Understanding the Semantic Web and Patterns, The Semantic Web Application, Use of Data Warehouse, RDF (Resource Description) and Graphs and SPARQL (Protocol and RDF Query Language).

REFERENCE

1. Karin Koogan Breitman. Web Semântica – A Internet do Futuro - Rio de Janeiro : LTC, 2005. P. 1-50.

2. Dean Allemang, James Hendler. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and QWL. Morgan Kaufmann Publishers, 2011. P. 1-20.

3. Webconferência link, Web Semântica, Professor Rafael Guem Murakami, 2023. Disponínvel em: https://animaeducacao.zoom.us/rec/share/GfFE_Lg-Y_NQNIHoC0lK5uWIxlTbQWG3_JYxl295l84efqz85-kWPI4aSwH84m0z.3Q1M7wHXUBf-TcqH

4. Web Semântica, material de aula, Anhembi Morumbi Universidade, EAD Laureate. Professor Rafael Guem Murakami, 2023.

5. EdUcational Curriculum for the usage of Linked Data, disponível em: EUCLID | EdUcational Curriculum for the usage of LInked Data (euclid-project.eu)

6. Vista do Linked data em bibliotecas, disponível em: Vista do Linked data em bibliotecas (febab.org.br)

7. Guia Web Semântica, disponível em: Guia_Web_Semantica.pdf (nic.br).

8. W3C Brasil - World Wide Web Consortium Escritório Brasil, disponível em: W3C Brasil - World Wide Web Consortium Escritório Brasil

9. 614117317 pdf – core.ac.uk, disponível em: 61417317.pdf (core.ac.uk)

10. Building data warehouse with semantic web data – Science Direct, disponível em: Building data warehouses with semantic web data - ScienceDirect

11. Criação de padrões na web semântica: perspectivas e desafios (redalyc.org), disponível em: Criação de padrões na web semântica: perspectivas e desafios (redalyc.org)

GLOSSARY

Blockchain – Distributed ledger with growing lists of records that are secured

Cryptography – Secure communication

Extrinsic – That comes from outside

Hyperlinks – Digital reference to data that the user can follow

Intrinsic – Internal

Artificial Intelligence – It is the intelligence demonstrated by machines when performing complex tasks associated with intelligent beings

Ontology – Structured set of terms and concepts that represents a knowledge about the world

Web Semantic – That studies the meaning of words, related words

Typed Links – words related

INDEX

A

Abstract, 7

C

Conclusion, 22

D

Development, 12

Glossary, 24

L

List of abbreviations and acronyms, 10

List of Ilustration, 8

List de Tables, 9

I

Introduction, 12

R

Reference, 23

S

Summary, 11