






## SEMANTIC WEB TECHNOLOGIES FOR THE INFORMATION RETRIEVAL ON WIKIDATA

TECNOLOGIAS DA WEB SEMÂNTICA PARA A RECUPERAÇÃO DA INFORMAÇÃO NO WIKIDATA

<sup>1</sup>Larissa Pavarini da Luz   
Universidade Estadual Paulista  
São Paulo, SP – Brazil

<sup>2</sup>Caio Saraiva Coneglian   
Universidade Estadual Paulista  
São Paulo, SP – Brazil

<sup>3</sup>José Eduardo Santarem Segundo   
Universidade Estadual Paulista  
São Paulo, SP – Brazil

### Correspondence

<sup>1</sup>E-mail: [larissapavarinidaluz@gmail.com](mailto:larissapavarinidaluz@gmail.com)

**Submitted:** 27/02/2018

**Accepted:** 26/10/2018

**Published:** 05/11/2018

Antiplagiarism Check  




**JITA:** IL. Semantic Web

e-Location ID: 019003

**ABSTRACT**

Information Retrieval is responsible for the storage and automatic retrieval of information, and these documents may consist of texts, web pages, audio, video, images, graphics and figures. Information Retrieval techniques have gained importance with the growth of the Web, because the unlimited amount of information can express the most diverse forms and levels of quality to what is expected. With this in mind, the present work studies methods and technologies capable of retrieving this information, focusing on searching structured databases called Linked Data, but specifically on the Wikidata project, a database structured using Semantic Web concepts, which brings together the knowledge from Wikipedia. Seeking to understand how this information retrieval is done in the Wikidata project, this research has the objective of presenting the media that Wikidata provides to RI and how they use the principles of the Semantic Web. The methodology used was an exploratory study based on the research and applied, since tests were done in the database of Wikidata. As a result, the characteristics of the various forms of data access and retrieval were identified, tracing the correlations between each of these forms, with the theoretical framework of the Semantic Web and Information Retrieval. It was concluded that Wikidata stands as a solid database, with a large volume of contents, quite current, that has a series of recovery mechanisms, capable of serving the most diverse applications on the Web, because these mechanisms are built with different technologies and configurations.

**KEYWORDS**

Semantic web. Information retrieval. Linked data. Wikidata.

**RESUMO**

A Recuperação da Informação é responsável pelo armazenamento e pela recuperação automática de informação, podendo estes documentos ser constituídos por textos, páginas Web, áudio, vídeo, imagens, gráficos e figuras. Técnicas de Recuperação de Informação ganharam importância com o crescimento da Web, pois a quantidade ilimitada de informação pode expressar as mais diversas formas e níveis de qualidade ao que se espera. Pensando nisso o presente trabalho estuda métodos e tecnologias capazes de recuperar essas informações, dando enfoque a buscar em bases de dados estruturadas chamadas Linked Data, mas especificamente no Wikidata, uma base de dados estruturada utilizando conceitos da Web Semântica, que reúne conhecimentos da Wikipédia. Buscando compreender como é feita essa recuperação da informação no projeto Wikidata, esta pesquisa tem como objetivo apresentar os meios que o Wikidata fornece para a RI e como eles usam os princípios da Web Semântica. A metodologia utilizada foi um estudo exploratório com embasamento para a pesquisa e aplicada, uma vez que testes foram feitos na base de dados do Wikidata. Como resultados, identificou-se características das diversas formas de acesso e de recuperação dos dados, traçando correlações existentes entre cada uma destas formas, com o arcabouço teórico da Web Semântica e da Recuperação da Informação. Concluiu-se que o Wikidata se coloca como uma base de dados sólida, com um grande volume de conteúdo que possui uma série de mecanismos de recuperação, capazes de atender às mais diversas aplicações existentes na Web, devido a estes mecanismos serem construídos com distintas tecnologias e configurações.

**PALAVRAS-CHAVE**

Web semântica. Recuperação da informação. Linked data. Wikidata.

## 1 Introduction

Regarding the Web, some services have a role of gathering large amounts of information from various domains, such as Wikipedia. With the consolidation of the Semantic Web, through initiatives such as Linked Data, these services dealing with general domains are being converted into databases following the principles of linked data. An exponent of this context is Wikidata, which gathers mostly information from Wikipedia.

Wikidata was created with the purpose of gathering structured information contemplating techniques capable of extracting the data of Web pages automatically. This feature has made the amount of information present on Wikidata grow very rapidly and is currently one of the main sources of structured data on the Web.

In this way, the amount of data that a base such as Wikidata gathers, is extremely voluminous, bringing great challenges to Information Recovery in these environments. This has prompted Wikidata to create a series of mechanisms for its information to be extracted and localized. In order to understand such mechanisms, the present paper questions: What are the means that Wikidata provides for Information Recovery, as well as the way in which such means use the principles of the Semantic Web?

In order to respond to this problem, the present work aims to analyze the modes of Information Recovery provided by Wikidata, identifying how the use of Semantic Web principles occurs within this service, and analyze how such principles make Information Recovery more contextualized according to the contents of this database. In this way, we try to analyze one of the main databases in the Web and Semantic Web, verifying if Wikidata follows the principles discussed by Semantic Web designers, Berners-Lee, Hendler and Lassila (2001), in which there is a need to be able to understand with greater precision the domain in which the contents are inserted.

As methodology, an exploratory study was carried out, searching the literature for a basis for the research, and that applied, tests were done exploring the database of Wikidata. It was identified that Wikidata is a rich data base, which is structured in formats compatible with the Semantic Web, and in addition, it was possible to verify that the ways of retrieving the data allow users to with a wide range of options.

In this way, it is possible to define the methodological content as follows: The bibliographical research was done through bibliographical surveys in national and international databases, exploring the themes: Information Recovery, Semantic Web, Linked Data and Wikidata.

The applied action was realized from the direct and participative observation in the site of the Wikidata, carrying out proofs of concepts from the information exposed by the service and the forms of access to the data.

In this perspective, this work analyzed the modes of Information Recovery provided by Wikidata, aiming to identify how the use of the principles of the Semantic Web within this service occurs, and how these principles make the Information Recovery more contextualized according to the contents contained in the database of data. With the identification of the forms of access to the data, as well as the accomplishment of actions to identify the functioning of such forms, it was possible to trace the correlations between each form of access, with theories about Information Recovery and the Semantic Web.

## 2 Semantic Web and Linked Data

The evolution in Information and Communication Technologies has been providing changes in the Web, changing the paradigms followed by this informational environment. Since its inception in 1989, the Web has been constantly changing, especially linked to the evolution of both technologies and the usefulness that users have found from their tools.

In this perspective, one of the main benefits of the Web deals with the knowledge that the data can provide, that pass through a better vision of the people and the institutions. In view of the above, the emergence of the Semantic Web occurred naturally, due to the need to have a better understanding of this information, which at first was incomprehensible to the computational mechanisms.

This scenario is exposed by Berners-Lee, Hendler and Lassila (2001, p.2, our translation), who claim that the Semantic Web is a: "[...] extension of the current Web, where information has a meaning clear and well defined, enabling better interaction between computers and people. "

Thus, the Semantic Web has emerged as a solution to allow a better understanding of the information available on the Web by the mechanisms, in order to provide bases for the creation of smarter applications and a better understanding of the domain in which data is found. Berners-Lee, Hendler and Lassila (2001) discuss this issue by pointing out a future scenario at that time, but in which computational agents would be able to facilitate the user's life by being able to extract, interoperate and understand data from various sources.

The Semantic Web proposal has been evolving from year to year, creating more consistent concepts and technologies capable of supporting applications that can accurately

explore databases, understanding the context of such information. This phase in which use applications are built and used in mass is being called materialization of the Semantic Web, process that has Linked Data as main application.

Linked Data has in principle to insert in a structure compatible with the Semantic Web, meaning to the data. This initiative was proposed by Berners-Lee in 2006, presenting some guidelines for the creation of databases following norms that facilitate the localization and the insertion of meaning in these data. Based on this proposal, a number of linked and open databases were created; following the proposal of Linked Data, using the technologies recommended by the Semantic Web, in particular DBpedia, which gathered information mainly from Wikipedia.

The consequences of this evolution of the Web, especially with regard to the Semantic Web and Linked Data, have promoted profound changes in several fields of study, especially Information Recovery. Seeking to point out these changes, as well as the correlations that the theoretical framework of this context has with the concepts and technologies of the Semantic Web, the following are concepts about Semantic Web Technologies in Information Recovery.

### **3 Semantic Web Technologies in the Information Recovery**

Due to the considerable and exponential increase in information stored and available for access, Sant'Ana (2008, p.145) stated that the adoption of information and communication technologies for the transmission of this information to the user is extremely relevant, especially in refers to Information Recovery (IR) processes.

Thus IR is classified as a solution to the problem of the informational explosion identified by Bush in 1945 as the irrepressible growth in the amount of data generated. Over the decades, this information explosion has become more potent, especially through the mass use of technologies.

The actions may be percebido, the release of the systems of information and communications data in the biggest activities, which based on a creation and use of data in formats digitais. Conseqüentemente, tornou-se mais difícil recuperar as informações de modo que as mídias digitais, além de criar uma multiplicidade de informações na Web, que possibilitaram a criação de computadores e os consumidores de informações, tornando a maioria dos ambientes digitais quase que totalmente integrada. geração de dados mais expressiva.

The difficulties to locate and retrieve data through this information medium, a Web, were launched with the creation of the Semantic Web, which are based on Santarém Segundo, Souza and Coneglian (2015) aimed at helping and improving the user's life, allowing that all information is searched for and is being interconnected and with explicit meaning.

In this sense, it is worth noting that the Semantic Web is not only related to the format of the content of a resource, but also to the way this content will be made available and will interact with other resources on the Web, that is, besides the human user being able to search by machines would be able to provide the correct return of the meaning that such data has.

In order to make the Semantic Web, and consequently to improve IR, a series of technologies have been developed, such as XML (eXtensible Markup Language), RDF (Resource Description Framework), RDF Scheme, OWL (Web Ontology Language, SPARQL (SPARQL Protocol and RDF Query Language) among other concepts that are described by W3C (W3C, 2017), which will be described in the sequence.

Created in 1998, the language eXtensible Markup Language (XML), an extensible markup language, is the language W3C (W3C, 2017) recommended to describe the metadata that a document contains, and its main goal is to create a unique infrastructure for several languages.

Berners-Lee, Hendler and Lassila (2001) made the whole proposition of the Semantic Web architecture based on XML, aiming to generate a standard of assimilation in the electronic document exchange, in a textual, simple, structured, extensible, flexible, semantically rich and with adequate security. Thus, the XML language allowed and allows the creation of concise structures in the documents contained in the Web.

It is worth mentioning that the Semantic Web has a working structure, syntax, that uses the Universal Resource Identifier (URI) to represent the data. The identification of their resources is done in a unique way, as well as their relationships, using the URIs to name them.

Another central technology for the Semantic Web is the Resource Description Framework (RDF), a declarative language that has become a standard recommended by the W3C in 2004. RDF represents the data in the form of statements about properties and relationships between resources, which can be virtually any object in the real world.

The RDF features focus on being domain independent and consists of a triple: subject, predicate, and object. The formation of a triple is due to the combination of a resource, a property, and a value for the ownership of a resource. (DZIEKANIAK; KIRINUS, 2004).

Other relevant features of RDF can be listed as: seeking a primitive representation for a larger creation, has the purpose of semantic enrichment, and provides the ability to communicate in a transparent way, or as close to it, in relation to semantics for metadata, facilitating the recovery of information on the web. RDF does not provide the necessary subsidies to be considered an ontology language, such as the Web Ontology Language (OWL) language.

The OWL has more flexibility in expressing meanings and semantics compared to XML / RDF and RDF Schema, which is designed to be used for the use of the DARPA agent markup language (DAML) + Ontology Inference Layer (OIL) in applications that need to process the content of information, rather than just displaying the information.

OWL is considered the core of Semantics (W3C, 2015), being a Web language based on RDF, and defined as a language for instantiation of ontologies. This language can formalize a domain by the definition of classes and properties, as well as define individuals and affirmations about them, able to specify how to derive logical sequences, that is, facts that are not present in the ontology, but which are bound by semantics.

The data recovery in RDF and OWL occur in the Semantic Web, through the SPARQL Protocol and RDF Query Language (SPARQL), which according to W3C (2008) constitutes the first layer of the Semantic Web, is classified as being a query language and a protocol capable of retrieving and manipulating data in RDF format, allowing the recovery of structured and semi-structured data values, exploiting data when querying unknown relations, performing complex joins of different data sets in a single, simple query.

Starting from the concepts presented, the following section presents the results and the discussions of the present work.

## 4 Results and Discussions

Data bases that follow the principles of Linked Data are usually a very comprehensive source of information, containing a high level of contextualization of the information contained therein. In this way, the use of these databases in several other applications can lead to an improvement in the level of understanding that the computational agents have of the content available on the Web.

This scenario meets the ideals presented by Berners-Lee, Hendler and Lassila in 2001, when they proposed the Semantic Web. However, the creation of interfaces capable of relating

and retrieving the data of these datasets is essential for the popularization of the use of these tools, especially with regard to the application in a practical and usual way.

In order to argue such questions, it is necessary that the theoretical bases for Information Recovery are required and discussed, aiming to identify how, from the need to access the data of these datasets can present an efficient Information Recovery, following the principles of the Web Semantics.

In the context of Linked Data, Wikidata is gaining notoriety for the amount of information it presents, as well as for the processes that this environment has, aiming at having a sustainable growth of its data, which seeks to identify incorrect data, allowing the correction. Another point of prominence of Wikidata is the relationship between this service and Wikipedia, allowing the automatic extraction of information from this second environment.

Thus, from the point of view of Information Recovery, identifying and analyzing the search forms that exist within Wikidata is fundamental to understanding the extent that this platform can obtain within the Web as a whole. It is worth noting that the Wikidata Semantic Web base allows a new perspective of communication with other Web services, but it should be analyzed if the principles of the Semantic Web are being used.

Wikidata points out ways that information can be sought by dividing it into two main groups (WIKIDATA, 2016): Access to item data and access to evictions. The division in these two classifications is made by the principle of data access, where one occurs centered on the location of specific items, and the second allows the user to access the entire set of data.

The named form of data access by items has several subdivisions, due to the way each one of them allows to recover and to access the data. Thus, the sequence seeks to trace the existing correlations between each of the existing forms with the theoretical framework of Information Recovery and Semantic Web. These are the divisions: the linked data interface, the MediaWiki API, the Wikidata query, the SPARQL Endpoint, and the Robots.

The linked data interface is a web page where the user can view the information and relationships that a particular item has. This interface can be accessed by the URI itself of a resource, presented in a visual interface and understandable to people the contents expressed in the database, of that specific item.

When accessing the URI, the system itself redirects to the Wiki page, allowing modifications and insertions in the contents available. The possibilities of this page include one of the principles of Wikidata, which is to be a collaborative linked database, which allows users to make changes to the content. It is worth noting that unlike what occurs when a change is



made on Wikipedia, in which the user changes the Web page, in Wikidata the user will follow the structures of the Semantic Web to make the changes, changing the existing relations themselves, being able to correct errors and non-conformities in existing connections.

From the standpoint of the Semantic Web, this interface allows human users to be able to visualize in a simpler way the information contained in the database. Another point to highlight that brings benefits, from the point of view of data semantics, is the possibility of changes in the data, as previously reported. This issue allows for a more constant update of the data, avoiding that Wikidata becomes obsolete. This recurring problem in other databases of Linked Data is solved at this point, along with other measures, such as the use of robots, which will be addressed in the course of the text.

For the Information Recovery, this visual interface contemplates a clear and representative visual form, being a final screen of information recovery satisfactory to indicate the data of a resource.

Figure 1 presents a linked data interface screen, featuring Tim Berners-Lee, a computer scientist and founder of the Web and Semantic Web.



**Figure 1.** Tim Berners-Lee feature screen

Source: Prepared by the authors from the consultation page <http://www.wikidata.org/entity/Q80> (Accessed on: January 23, 2018)

In this figure it is possible to identify the information that the resource has, besides demonstrating how the users make the changes in the contents, being conditioned to insert or to change the data following the RDF, triple standard. These relationships can be visualized by the highlights realized in red in figure 1.

A second way to access data is through the MediaWiki API, which allows external applications to query the data by accessing a Wikidata server. In the definition of this API, there are a series of parameters that allow the query, defining the standards that should be used in the search.

The search through the MediaWiki API has an important character for Information Recovery, since it is intended for computational applications that want to build their own standards to consult the structured data of Wikidata. An in-depth analysis of the parameters of this API indicates a consistency with the major existing search models within the Web.

It is worth mentioning that the API allows a range of options ranging from login and logout in the system, through questions related to the formats of the data, even to options related to the search itself. In the scope of this research, the analysis of the possibilities offered regarding the search, more specifically, the search about entities, that accurately reflects the semantic structure in which the data are presented, is centered.

Table 1 presents the main parameters for performing an entity search within Wikidata, using the MediaWiki API.

**Table 1.** Basic parameters for defining a MediaWiki API search

<b>Parameter</b>	<b>Possible values</b>	<b>Comment</b>
<i>Action</i>	All allowed actions, in the case for entity search, use the value "wbsearchentities"	This parameter defines which action will be taken within the API.
<i>Search</i>	You should search for a textual value (string) containing the elements you want to search.	This parameter should be used when the action is "wbsearchentities", returning the entities that possibly meet the search performed.
<i>language</i>	Use language code, such as pt-br (for Brazilian Portuguese) or en (for English)	It will define the language that the results obtained

<i>format</i>	json, jsonfm, none, php, phpfm, rawfm, xml, xmlfm	Helps to define the most appropriate format for each application
---------------	---	--

Source: Prepared by authors from API queries

From these parameters indicated in table 1, it is possible to perform a query, retrieving a series of entities. In this way, a search was constructed, following the mentioned parameters that obtained the URL to make the query, the following content: "https://www.wikidata.org/w/api.php?action=wbsearchentities&search=Tim%20Berners-Lee&language=en". It is worth noting that the only parameter not used in the charts was the format, since only an easy-to-understand visual response was sought; in addition, other parameters could be used to give a greater level of detail in the results.

This query built searches for entities referring to "Tim Berners-Lee" in the English language. The results obtained are presented in figure 2, in which are listed the entities that have in their name the text informed as search.

```

{
  "searchinfo": {
    "search": "Tim Berners-Lee"
  },
  "search": [
    {
      "id": "Q80",
      "concepturi": "http://www.wikidata.org/entity/Q80",
      "url": "//www.wikidata.org/wiki/Q80",
      "title": "Q80",
      "pageid": 139,
      "label": "Tim Berners-Lee",
      "description": "British computer scientist",
      "match": {
        "type": "label",
        "language": "en",
        "text": "Tim Berners-Lee"
      }
    },
    {
      "id": "Q22991023",
      "concepturi": "http://www.wikidata.org/entity/Q22991023",
      "url": "//www.wikidata.org/wiki/Q22991023",
      "title": "Q22991023",
      "pageid": 25007543,
      "label": "Tim Berners-Lee: A Magna Carta for the web",
      "description": "TED2014",
      "match": {
        "type": "label",
        "language": "en",
        "text": "Tim Berners-Lee: A Magna Carta for the web"
      }
    }
  ]
}

```

**Figure 2.** Fragment obtained from the "Tim Berners-Lee" query in the MediaWiki API

Source: Prepared by the authors, from the consultation at:  
<<https://www.wikidata.org/w/api.php?action=wbsearchentities&search=Tim%20Berners-Lee&language=en>>.  
Accessed on: 25 Jan. 2017.

In the results obtained with the query, presented in figure 2, it is possible to identify that two entities were recovered, all containing "Tim Berners-Lee" in the name, inserting the computer scientist in the first position and a book in the sequence. This fact is interesting because it allows each application, when using the MediaWiki API, to delimit how the results will be presented to users, or the use that will be made of each entity.

Another point to emphasize is that this type of search returns only an indication of the localized entity, without presenting the existing relations between the entities and their properties. When analyzed from the perspective of the Semantic Web, this API contemplates fundamental elements that make the interoperability and the communication between computational agents and contents created by people, especially when delineating consistent identification variables like the URI that points an address and a name for that entity specific.

The third mode presented to access the data is through the Wikidata Query. This tool has the purpose of allowing queries to be built with a high level of complexity, aiming to relate the data and to explore the entities of Wikidata.

This form of access is more convergent with the main theories of Information Recovery, since it facilitates the creation of queries using models from this field of study. Among them, the possibility of using logical operators such as the AND and the OR is highlighted, allowing the construction of Boolean queries to retrieve the information more adherent to the informational needs.

Another highlight of this tool is the interface that allows searches using the RDF triple property structures, as well as giving the possibility to search on the instances and subclasses that organize the information within that database. The functions allowed by this mechanism, make explicit a direct relation with the concepts of the Semantic Web, since the functionalities of this proposal are adherent, especially to that which corresponds to the RDF triples.

The construction of queries in this mechanism presents some difficulties, due to the complexity made possible by the tool; however, it is able to precisely meet the purpose of consulting the data of Wikidata, inserting a series of elements that make this process better. The options available for the construction of the queries can be analyzed in the page of the WikidataQuery API (2017) ([https://wdq.wmflabs.org/api\\_documentation.html](https://wdq.wmflabs.org/api_documentation.html))

In order to demonstrate how this type of query works, Figure 3 represents the standard query construction screen, which, together with the recommendations for using this tool, allows the construction of queries on Wikidata.

## Wikidata Query editor

### Create and edit queries

Check out the [API documentation](#) for details on how to write a query, or use this handy editor!

TREE[30][150][17, 131] AND CLAIM[138:676555]

TREE ▾	AND CLAIM ▾	
Root items	Prop/Item/Query	
United States of America [Q30]	named after [P138]	: Francis of Assisi [Q676555]
Forward		
contains administrative territorial entity [P150]		
Reverse		
country [P17]		
in the administrative territorial entity [P131]		

Show the results for the current query in : [Autolist](#)

### Examples

- [Places in the U.S. that are named after Francis of Assisi](#)
- [All items in the taxonomy of the Komodo dragon](#)
- [All animals on Wikidata](#)
- [Bridges in Germany](#)
- [Bridges across the Danube](#) (alternate language labels example: [German](#))
- [Items with VIAF string "64192849"](#)
- [People who were born 1924-1925, and died 2012-2013](#)
- [Items 15km around the center of Cambridge, UK](#)

### Code

The query editor above is embeddable in your own web tool! Just check out the source of this web page, or the [BitBucket](#) source!

**Figure 3.** Example of using the Wikidata query

Source: Prepared by the authors

The example shown in figure 3, contemplates an example search given by the Wikidata Query itself, which presents the places of the United States whose name contains Francisco de Assis. In this example it is possible to identify the occurrence of the Boolean AND logic, by relating two search information, as well as the use of the RDF properties to search for the information, in case the "named after" property was used to trace the relationship.

The results obtained with this query are listed by displaying the entities that match the search performed, similarly to that demonstrated using the MediaWiki API.

Another way to access Wikidata data is through the use of SPARQL Endpoint. This form is most integrated with Semantic Web proposals, because it is centered on one of the proposed basic technologies, SPARQL, responsible for being the data query language in RDF and OWL. The SPARQL Endpoint is an open environment for querying data from a database, which through SPARQL allows users, human or otherwise, to access such data.

The SPARQL Endpoint of Wikidata represents the way to query the data that takes advantage of the potential of the Semantic Web, being natively integrated to the other technologies of Semantic Web and Linked Data. In this way, it becomes possible to more accurately explore the context of the data, while allowing the construction of inferences and axioms capable of expanding the search.

In addition, SPARQL supports integration with other platforms that also have SPARQL Endpoint. In this way, the RDF language together with SPARQL favors interoperability between available data information in formats compatible with Linked Data, favoring the reuse and exchange of data.

From the perspective of Information Recovery, SPARQL is able to provide a tool that exploits the data, without following the classic paradigms of this field of study. However, it is identified that SPARQL acts in a step of Information Recovery, being linked to the data localization processes and aiming the relationship between the information.

As a demonstration and use example, a search was performed through the SPARQL Endpoint, which shows all the people who were born in the city of Marília, Brazil.

This search expresses one of the possibilities that the work with SPARQL allows, in that it is possible to draw relationships, searching for information of a high level of specificity. In short, locating all the personalities that were born in the city of Marília, if you do not use a tool with this, it is not easy to get to the result, because sometimes it is necessary to go through several sites, locating in some spaces where such information can be found. SPARQL makes this process simpler as it uses existing relationships to promote that search.

Figure 4 demonstrates the results achieved with the search for people born in Marília, in the SPARQL Endpoint environment. It is noteworthy that the search was limited to the first five people recovered, in addition to the property P19 has as a characteristic to present the place of birth of a person.

The screenshot shows the Wikidata Query interface. At the top, there is a header with the Wikidata logo and the text 'Wikidata Query'. Below this is a text area containing a SPARQL query:

```

1 PREFIX p: <http://www.wikidata.org/prop/statement/>
2 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3 SELECT ?s
4 WHERE {
5   ?s p:P19 ?o.
6   ?o rdfs:label "Marília"@pt-br.
7 }
8 LIMIT 5
    
```

Below the query area, there is a status bar that says 'Press [CTRL-SPACE] to activate auto completion.' and 'Data updated 2 minutes ago'. Underneath, there are buttons for 'Run', 'Clear', and a status indicator '5 Results in 341 ms'. There are also dropdown menus for 'Display', 'Download', and 'Link'. The results section shows a list of five Wikidata statement IDs:

- wd:statement/Q47314-908B1DDB-FA94-4032-A3E9-EA83D4BBEA85
- wd:statement/Q569094-BA349C5C-8397-4AA9-A7EE-98D749D1A347
- wd:statement/Q3869199-B7935AEF-471D-4F82-AD9F-C1F59A4D5A58
- wd:statement/Q3941032-F6A16EB9-D463-42A2-A34D-202039B97403
- wd:statement/Q6298236-12743294-7A47-4E05-9F8C-0EF6A4352FC3

**Figure 4.** Search for people born in the city of Marília  
 Source: Prepared by authors from query in the SPARQL Endpoint of Wikidata.

The results obtained with the query presented are entities that present a series of other properties. Such feature allows such data to be explored in depth only with the use of the SPARQL language, thus allowing a wide range of options, which can be used to relate certain information, as well as in computer programs that use data on the Web, such as example of decision-making systems, among several other scenarios in which the use of such data may be useful.

The last form of access to the data by item, are the so-called robots, whose function is to perform changes and insertions of information automatically. These robots are built by the users themselves who aim to collaborate with the proposal, however, the ease and speed with which these robots are able to change the contents of Wikidata, led the community to draw up a policy for the construction of robots. (WIKIDATA, 2017b).

In short, this policy will define fundamental questions about changes and insertions of content, together putting together a process of approval of these modifications through some robots called robots administrators. In addition, there are a number of requirements that robots must have to access the data from Wikidata.

The use of these robots is one of the main highlights of Wikidata, compared to other databases of Linked Data, due to the speed with which the data can be inserted or changed, keeping always updated.

Finally, the other form of data access is Access to evictions, or access to the database dumps. This possibility meets the guidelines of Linked Data that propagates that the data must be made available openly, allowing each user to make the best possible use of their conditions.

Downloading this raw data can occur in a variety of formats, such as JSON (JavaScript Object Notatio), XML and RDF, allowing the user to identify the format that best meets their needs. As reported, from the standpoint of Semantic Web, this access is quite interesting, so that new applications can be constructed, and if none of the presented data access options are useful to an application, it can create more adequate forms of access from of raw data.

Within the concept of Information Recovery, the option of downloading the raw data leaves the application in the hands of the best way to build the processes to recover the data. However, by allowing several options for data availability, it facilitates the construction of efficient forms of recovery, especially for the versatility of this option.

All of the options presented are centered on the Semantic Web, as well as enabling ways to improve Information Recovery in a context of changes, both in the way information is entered on the Web, and in relation to how applications are consuming the data . Semantic Web and Linked Data technologies are helping a new way of organizing and retrieving data within the Web, in order to insert more information regarding the context of such data.

In this way, Wikidata presents a series of options that are based heavily on theories and technologies of both the Semantic Web and Information Recovery, being a very updated platform, being adequate the current needs of the Web.

The headings of the sections / subdivisions should be brief and clear. The text of the article should be structured preferentially contemplating the following items: introduction, method, results and final considerations. Acronyms and abbreviations must be enclosed in parentheses and preceded by their full meaning when first used in the text.



## 5 Considerations and Implications

The influence of concepts and technologies of Semantic Web is occurring in several fields and fields of study, especially to all those that are directly or indirectly related to Web. This influence has as principle to make the relationship of people with computational mechanisms more natural, since it allows an integration and an understanding of the contents made available on the Web by these two actors.

In this perspective, the creation of content by the users themselves, which has on Wikipedia its main exponent, provoked a revolution in the Web, since the amount of information made available for queries grew as the number of users increased. However, such information was mostly understandable only by people, with computational agents, information recovery systems, and search engines only syntactically accessing the contents of these platforms.

In order to solve this issue, the Semantic Web was proposed with the objective of "intelligent" access to heterogeneous data, bringing a leap in efficiency in the way information is searched and displayed. This efficiency is due to the fact that the information becomes meaningful, allowing the understanding of the contents by human beings, but also by the computational systems.

In order for all these concepts to really be useful for people, inherent users of technologies, Information Recovery systems play a fundamental role in this context, since it is such systems that make the information tangible to the user, either by promoting the search done in documents, or in the search for metadata. In this way, it is necessary to approach Information Recovery and the Semantic Web, when analyzing initiatives that are inspired by the Semantic Web, whose purpose is to gather data from a wide range of domains.

In this sense, the present study deepened the understanding about Wikidata, analyzing how this application based on the principles of Semantic Web and Linked Data, promotes the Information Recovery of its data.

The results obtained indicated several ways for users, both human and non-human, to access and use linked data from Wikidata. The analyzes carried out in each of the ways of accessing the data showed particularities, both in terms of the use of Semantic Web concepts and in the methods of Promoted Information Recovery.

These differences demonstrate that Wikidata is prone to encompass an extensive number of applications that wish to use their database as an informational source of Linked Data for a variety of purposes. Another important point is the robustness of the data search

solutions, where there is a focus on maintaining the data structure in the source formats, in the case of RDF.

All of these ways differentiate ways of retrieving data, without losing the origin based on the Semantic Web and Linked Data formats. It should be noted that SPARQL is the most readable language for retrieving data, but the option of accessing with more common web solutions, such as query APIs, makes it more comprehensive.

Therefore, Wikidata can be seen as a solution that is able to gather content that at the beginning was understandable only by people, originating mostly from Wikipedia, in formats understandable by machines, especially RDF and OWL. The work carried out also demonstrates that this database, Wikidata, can be quite prominent and important on the Web, because it has search engines and Information Recovery efficient and accessible by applications built from different configurations.

## References

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific american**, v. 284, n. 5, p. 28-37, 2001.

DZIEKANIAK, G.,V.; KIRINUS, J. B. Web Semântica. Semantic Web. **Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.**, Florianópolis, n.18, 2º sem. 2004. Available on: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2004v9n18p20>. Access on: 31 jan. 2018.

SANTAREM SEGUNDO, J. E.; SOUZA, J. O.; CONEGLIAN, C. S. Web semântica: introdução a recursos de visualização de dados em formatos gráficos. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., João Pessoa, PB. 2015. **Anais eletrônicos...** João Pessoa, PB: ANCIB, 2015. Available on: <http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/view/2780>. Access on: 2 feb. 2018.

SANT'ANA, R. C. G. A importância do papel do profissional da ciência da informação nos processos de recuperação de conteúdos digitais estruturados. *In*: GUIMARÃES, José Augusto Chaves; FUJITA, Mariângela Spotti Lopes (Org.). **Ensino e pesquisa em biblioteconomia no Brasil: a emergência de um novo olhar**. Marília: Cultura acadêmica, 2008. p. 145-154.

SILBERCHATZ, A.; KORTH, H. F.; SADARSHAN S. **Sistema de banco de dados**. 6º edição traduzida. Editora Elsevier. 2012.

WIKIDATA QUERY API. **Query construction**. 2017. Available on: [https://wdq.wmflabs.org/api\\_documentation.html](https://wdq.wmflabs.org/api_documentation.html). Access on: 27 jan. 2018

WIKIDATA. **Wikidata: robôs**. 2017b. Available on: <https://www.wikidata.org/wiki/Wikidata:Bots/pt-br>. Access on: 27 jan. 2018.

WIKIDATA. **Wikidata: acesso aos dados**. 2016. Available on: [https://www.wikidata.org/wiki/Wikidata:Data\\_access/pt-br](https://www.wikidata.org/wiki/Wikidata:Data_access/pt-br). Access on: 25 jan. 2018.

W3C. **Web semântica**. 2015. Available on: <http://www.w3c.br/Padroes/WebSemantica>. Access on: 31 jan. 2018.

W3C. **SPARQL query language for RDF**. 2008. Available on: <https://www.w3.org/TR/rdf-sparql-query/>. Access on: 31 jan. 2018

W3C. **OWL: web ontology language (OWL)**. 2012. Available on: <https://www.w3.org/OWL/>. Access on: 31 mar. 2017

W3C. **Extensible Markup Language**. 2015. Available on: <https://www.w3.org/XML/>. Access on: 30 jan. 2018.



This work is licensed under a Creative Commons Attribution 4.0 United States License.



This journal is published by the [University Library System](#) of the [University of Pittsburgh](#) as part of its [D-Scribe Digital Publishing Program](#) and is cosponsored by the [University of Pittsburgh Press](#).