UnBGOLD: UnB Government Open Linked Data

Semantic enrichment of open data tool

Luiz C. B. Martins
University of Brasilia (UnB)
Brasília - DF, Brasil
luizmartins@unb.br

Márcio C. Victorino
University of Brasilia (UnB)
Brasília - DF, Brasil
mcvictorino@uol.com.br

Maristela Holanda
University of Brasilia (UnB)
Brasília - DF, Brasil
mholanda@unb.br

George Ghinea
Brunel Univerity London
London, UK
george.ghinea@brunel.ac.uk

Tor-Morten Grønli
Kristiania University College
Oslo, Norway
tmg@kristiania.no

ABSTRACT

In accordance with current legislation designed to make public management more efficient and transparent, Brazilian Federal agencies have adhered to an open data publication policy, despite the challenge presented by datasets being published collectively rather than in isolation. Aiming to facilitate this process, this article presents the UnBGOLD, which adresses the need to connect the data in order to facilitate the publication of open semantically enhanced data. It is a tool that couples the architecture of open data publishing of the University of Brasilia and makes it possible to transform datasets into linked open data utilizing metadatas and ontologies in RDF formats, aside from making it possible for the data to be published automatically on the CKAN platform.

CCS CONCEPTS

• **Information Systems** → **Information retrieval**; • **scalability** → *Search engine indexing*;

KEYWORDS

Open Data, Linked Data, Ontology, UnBGOLD, RDF

ACM Reference Format:
Luiz C. B. Martins, Márcio C. Victorino, Maristela Holanda, George Ghinea, and Tor-Morten Grønli. 2018. UnBGOLD: UnB Government Open Linked Data : Semantic enrichment of open data tool. In *The 10th International Conference on Management of Digital EcoSystems (MEDES '18), September 25–28, 2018, Tokyo, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3281375.3281394

1 INTRODUCTION

In recent years, Brazil has invested in increasing active civil participation in government through initiatives that promote the transparency of governmental actions. The population can now monitor and supervise is government leaders and thus assist in suggestions and criticism aimed at promoting efficiency in public management.

MEDES '18, September 25–28, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-5622-0/18/09…$15.00
https://doi.org/10.1145/3281375.3281394

The data disclosure policy in Brazil arose out of the need to attend to the mandates of the Law 12.257/2011, known as the Law on Access of Information (LAI), sanctioned on November 18, 2011. LAI was created with the objective of regulating the constitutional rights of all citizens to have access to governmental information of public interest [6] and in this context, includes publication of open data established by decree Nº 8.777, of May 11 of 2016 [7] that meets the desire for active transparency, where the State takes the initiative to increase the transparency of management. To increase the visibility of data, the Federal Government created the Brazilian Open Data Portal (PBDA)[1], a tool that centralizes publicity and access to government data in raw format so that any citizen can use them in the way that suits them [11].

Currently several Brazilian government agencies publish their data on their own open data portals and are cataloged in the PBDA following the standards defined by the National Open Data Infrastructure (INDA)[1], However, these standards are not intended to create a link in these databut rather standardize the means of publication. For the most part, the data sets cataloged in the PBDA are structured in open format as *Comma-separated values* (CSV), *JavaScript Object Notation* (JSON) and *eXtensible Markup Language* (XML), being that to extract information from the intersection between different datasets it is necessary a process of *Extract, Transform andLoad* (ETL) is necessary to manually make the connection between them. In this context, this work presents the tool UnBGOLD, which aims to receive the datasets in the format currently published by the government organs and offer the option of enriching them through controlled vocabulary using metadata and ontologies that semantically represent it by transforming the open data into connected open data.

UnBGOLD is part of the architecture developed by the University of Brasilia to improve the quality of publishing its data. We opted for a coupled architecture that would allow other organs to use this tool to also publish their data, regardless of the source, besides offering an interface where it is possible to streamline the process through the automation of the publishing.

To present this tool, this article is divided into sections as follows. In Section 2, the themes and technologies that were used to implement this work are presented. In Section 3 we explain the UnB publishing architecture, the technologies and the operation of

[1] http://dados.gov.br/

the UnBGOLD tool, as well as presenting an example of semantic indexing of a dataset/set of data. Finally, in Section4 the conclusions and the impact of the use of the tool in increasing the quality of published open data are presented.

2 STATE OF THE ART

According to the definition of *Open Knowledge Foundation* (OKF)[2], data is considered open when it is published, able to be used and reused freely, reproduced or processed independent of special interests, [10] – the only restriction being the reference of the data's origin. For data to be categorized as Government Open Data (DAG), it must follow the principles of being primary, complete, current, accessible, machine processable, with non-discriminatory access, non-proprietary and licence free[12]. According to activist and researcher David Eaves[8], open data is characterized by the following laws:

- If the data cannot be found and indexed on the Web, it does not exist;
- If it is not open and available in a machine comprehensible format, it cannot be reused;
- If some legal device does not allow its replication, it is not useful.

Much of the data published on the web daily is done so in an uncontrolled and disorganized way. In this context Tim Berners-Lee[4] presented an evolved version of the Semantic Web *web* where the data would be published in a way that computational agents could read and extract information in a smart manner to aid humans in their daily lives. Recognizing that the implantation of the Semantic Web is very complex and given the steady increase of open data publication, Berners-Lee proposed a principle to categorize the level of openness of a single data, such that the quality of a data is inversely related to its isolation. In other words, the more capacity the data has to connect to other data, the higher the quality [3]. This index is known as "5 Stars Linked Data"(*5 Stars Linked Data*), where the more stars that are attributed to a data, the higher the quality. This 5 Star scale, with the characteristics of each classification is as follows:

(1) If the data is available online with open license independently of its format, it is considered a data with one star;
(2) If the data available is in some format considered structured (like XLS), it is a data with two stars;
(3) If it is available in a structured format non-proprietary like CSV, JSON ou XML, it has three stars;
(4) If it is possible to identify the data through an IRL and it is in conformity with the patterns established by the *World Wide Web Consortium* (W3C), (RDF and SPARQL), in a way that it is possible to direct publications, it is considered a data with four stars;
(5) If, after it fulfills the previous rules, the data is connected to other data in a way that it creates a connection logic, then it is considered a five star data – of the highest quality..

Figure 1 presents this five star index visually.

To describe the data on the web, an RDF model, recommended by W3C, is used. In this model, the data is described by an RDF's

[2]http://br.okfn.org

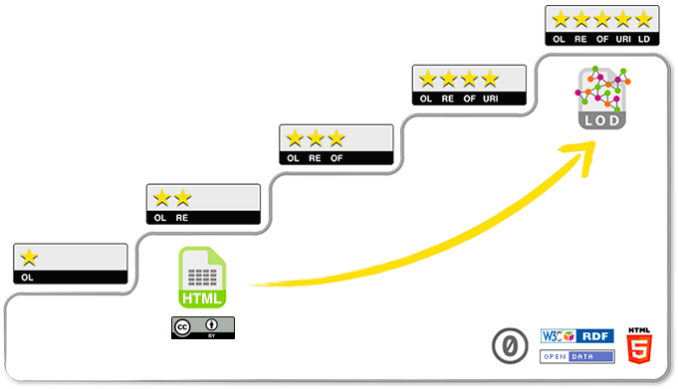

Figure 1: Five Star Scheme for Linked Open Data
Source: online http://5stardata.info/pt-BR/

triple: two entities called "Subject" and "Object" are related through a third one, called "Predicate", creating a form of a graph [5].

The Subject of the triple will always be a resource identified through an IRI while the Object can be a chain of characters that we call "Literal"", which is the explicit representation of the data, or a note to another RDF resource (through IRI)..

When defining an object as a resource, a connection is created, also with the triples, to which this resource is connected, increasing the link of the data. Subject and Object relate through a Predicate, which is a semantic representation that describes the characteristic of the link of the entities that must be made through the semantic representation by a controlled vocabulary utilizing already existing metadata and ontologies.

The triple RDFs can be represented in many ways. Figure 2 presents an example of a triple RDF described in XML and a graph generated by it.

The original RDF/XML document

```
1: <?xml version="1.0"?>
2: <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3:   xmlns:dc="http://purl.org/dc/elements/1.1/">
4:   <rdf:Description rdf:about="http://www.w3.org/">
5:     <dc:title>World Wide Web Consortium</dc:title>
6:   </rdf:Description>
7: </rdf:RDF>
8:
```

Graph of the data model

Figure 2: An Example of a Triple RDF
Source: online https://www.w3.org/RDF/Validator/

For the processing of those triple RDFs, the Apache Jena can be utilized. The Apache Jena is a free code for application development for Semantic Web and Connected Data for Java. Inside the *framework* many *Applications Programming Interface* (APIs) exist to manipulate and process RDF files – among them are SPARQL which is a research language in RDF and TDB files, which is a data base manager to store SPARQL queries and triple RDFs [9]. Figure 3 presents the architecture of the Apache Jena.

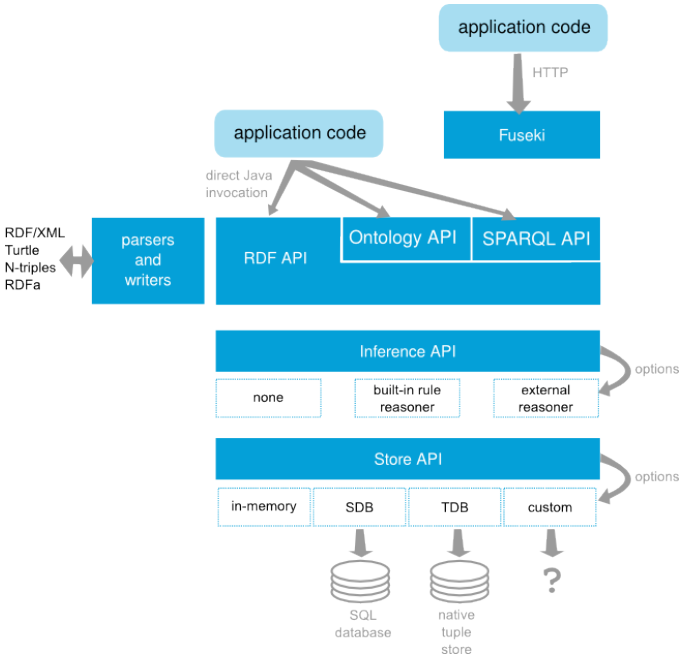

Figure 3: Architecture Apache Jena
Source: online http://jena.apache.org/getting_started/index.html

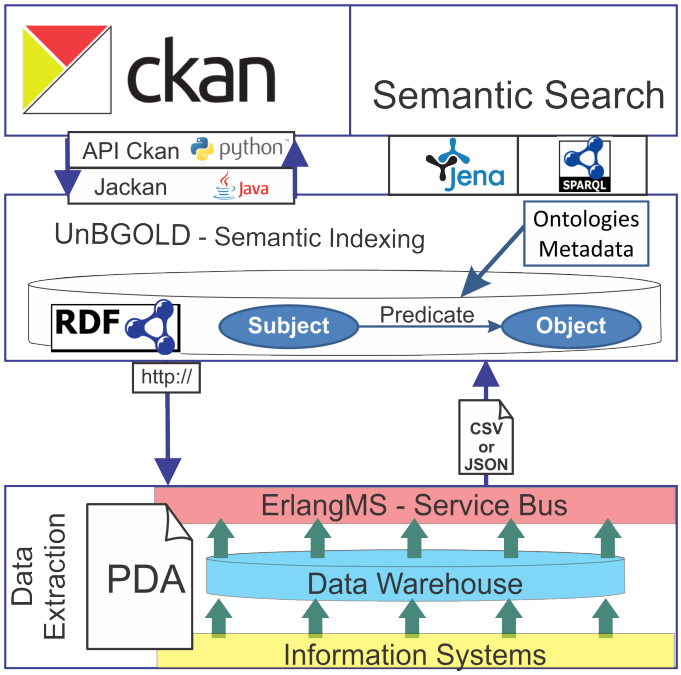

Figure 4: Open Data Publishing Architecture

3 UNBGOLD - UNB GOVERNMENT LINKED OPEN DATA

To increase the quality of data published by UnB, the Computer Center (CPD), in partnership with the Post-Graduation Project in Applied Computing (PPCA), proposed an architecture that facilitates the publishing of open data and that makes the semantic indexation of data possible. This architecture has three layers:

- **Data Extraction**: Through the collecting of data, importation to a textitData Warehouse and availability through a service;
- **Semantic Indexing**: Performed by the UnBGOLD tool (UnB - Governmental Linked Open Data) that offers an interface for semantic indexation of datasets utilizing metadatas and ontologies.
- **Data Publication**: Utilizing the API of the CKAN (Comprehensive Knowledge Archive Network).

Figure 4 presents the architecture of open data publication of UnB, such that the first layer concerns the data origin where a solution was implemented through a *Data Warehouse*, which already possesses the clean data originated from data banks of UnB's information system. The management of data solicitations and delivery is carried out by a barring of services ErlangMS[2] that receives a solicitation through a requisition via HTTP and returns the data in structured open format (CSV or JSON).

The last layer concerns the interface of the users' final consumption. UnB utilizes the solution CKAN[3], which is a web platform designed *Open Knowledge Foundation*(OKF)[4] for the publication and sharing of open data. In addition, a tool of semantic search is available on data already indexed. In the subsection **??** the functionalities of the UnBGOLD and how it communicates with the other layers is presented in detail.

3.1 Information Flux

The UnBGold is responsible for the communication between the extraction of data and its publication. Initially, a request is made via HTTP for the barring of services with the parameters defined by the Publication Agent, which will be presented in detail in subsection 3.3. The layer of extraction of data performs the recovery of data and returns the data conjunction so the semantic indexation can be made. Ultimately, according to the defined programming, the publication of data is performed, which can be done in three formats: JSON, CSV and RDF. The data can be published in one instance of the CKAN defined by the publication agent, in UnB's case, the official instance of publication is the university's open data portal [5].

3.2 Architecture

The UnBGOLD is a web application developed in Java language, that utilizes diverse technology, presented in Figure 5.

The users' view layer relies on the JavaServer Pages (JSP) to render the pages and the framework javascript AngularJS[6], performing the manipulation and communication with the control layer, aside from utilizing a customized version of the web responsive template

[3]https://ckan.org/
[4]https://okfn.org/
[5]http://dados.unb.br
[6]https://angularjs.org/

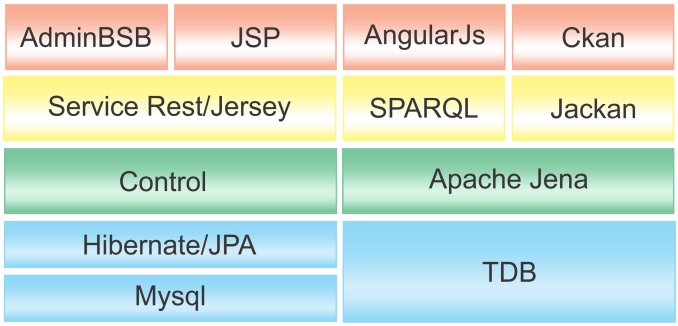

Figure 5: Architecture UnBGOLD

AdminBSB[7]. This layer is accessed by the Publication Agent, while the users consuming the published data utilize the CKAN platform.

Meanwhile, with regard to the persistence of the configuring and registering data, the application uses the data bank manager system MYSQL[8], being utilized the *framework* Hibernate and the API JPA for the object-relatonal mapping and managing the persistence of the data. The TDB[9], that is a component of the Apache Jena, is used for storing queries and triple RDFs for saving the already indexed data.

In the application control layer, we use the Jersey library to perform the Rest services management that feed the data to the view layer. In addition, the Apache Jena platform is used and is responsible for the manipulation of the triple RDFs. The tool provides an interface of semantic search that uses the SPARQL language to search through the triple RDFs.

Lastly, the UnBGOLD interface is used with the CKAN, which, through the services offered by the API of the CKAN, automizes the publication of open data and uses the Jackan[10] package that performs this interface. It is necessary that the Publication Agent informs which ones are the access keys to the CKAN instance in the agency, where the tool will perform the publication management.

3.3 Application

The main focus of the UnBGOLD application is to perform semantic indexing in data and to automate publishing. In this context, initially it is necessary to perform the registers that will be suitable as a base for the indexing. These registers are:

- **Publishing Entities**: Registering of the agencies that wish to utilize the UnBGOLD tool to aid the publishing;
- **Publicator Agents**: Users that will be responsible for indexing and publishing the data;
- **CKAN Instances**: Registering the CKAN instances that the UnBGOLD will communicate for the data publishing;
- **Vocabulary**: Registering of Ontologies and Metadatas that will be used by the Publication Agents to index the data.

The interface offered for the Publication Agents to publish their data is divided in four steps accessed sequentially. Those steps are: Publishing and Automation, Information about the data, Controlled Vocabulary and Semantic Indexing. Figure 6 presents that interface.

Figure 6: Publishing and Automation Step

In the example, a data conjunction referring to the majors at the UnB is used, and possess three fields: code, name and department, such that the reduced number of fields was utilized to make the explanation simpler.

Moreover, Figure 6, presents the Publication and Automation step, where the Publication Agent can opt to automate and configure the publishing. If they do not choose that option, it will have the possibility to only perform the download of the RDF file referring to the indexing. To configure the automation, the Publication Agent must inform which of CKAN instances previously registered wants to publish the data, the frequency of the publishing, that can be daily, weekly, bimonthly, semiannual, school semester and annual, as well as the time that it wants the data to be published and the format that this data will be published (CSV, JSON and/or RDF).

In the "Information About the Data" step, it is necessary to define the origin of the data through a URL where the file, which is already available, can be downloaded. It will also be possible to inform the parameters with which it will be utilized in the HTTP requisition. Those parameters can be fixed, leaving the desired value explicit on the parameter, or temporal, in which the Publication Agent must inform one of the temporal value options (day, week, month, semester, year...). The value of the parameters that will be sent, refer to the current moment, such that if the parameter is "daily", the parameter value will be the day that the request was made. The UnBGOLD will perform a connection test, accessing the data conjunction. If it works as expected, a table with the initial lines of the data conjunction for the Publication Agent conference will be available. In Figure 7, we present this screen, with the configuration of our example. The extraction data field is informed by the URL that contains the data collecting service and two parameters – one temporal and one fixed.

Next, the main information that identifies the data conjunction are informed. This data are standardized from the metadata defined

[7]https://github.com/gurayyarar/AdminBSBMaterialDesign
[8]https://www.mysql.com/
[9]https://jena.apache.org/documentation/tdb/index.html
[10]https://github.com/opendatatrentino/jackan

by the Government Open Data Publication Primer Federal[11]. This metadata can be obligatory or optional, given that some of the metadata are automatically generated.

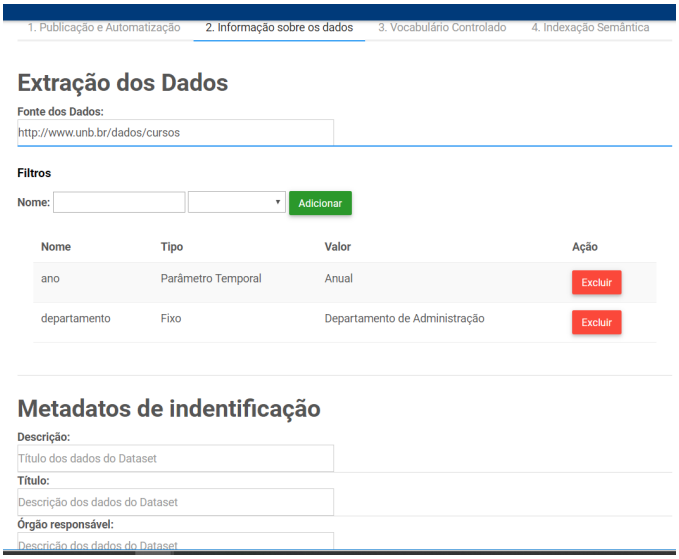

Figure 7: Information About the Data Step

In the Controlled Vocabulary step, a list of metadatas and ontologies previously registered is presented, as already explained in the beginning of this section. It is important to point out that the Publication Agent must know the data conjunction as well as the vocabulary that it wishes to use, thus, it will select the more adequate vocabulary to perform the semantic indexing of data. Figure 8 presents the interface, which is chosen by selecting the vocabulary conjunctions *Academic Institution Internal Structure Ontology* (AISSO) and D*CMI Metadata Terms* (DC) that match the example presented.

Figure 8: Controlled Vocabulary Step

Lastly, we present the semantic indexing of data screen. First, it is necessary to select which term of the vocabulary will define the data type being indexed. Afterwards, the Subject of the RDF is defined, given that it is necessary to define the IRI of the resource being used, which can be done through the combination of a text field with the data of a conjunction of the data field or a simple definition of a field that will be the IRI, as long as the resulting value is a URL and does not repeat itself. In the example in Figure 9, the IRI of the Subject is created from a URL informed by the Publication Agent and the complement is utilized in the "code" camp.

A conjunction of common data is a file with structured data organized in lines and columns. To indexes semantically the dataset; triple RDFs are created where the dataset line is identified with the Subject and the columns are the Object. Here it is necessary for the Publication Agent to define the vocabulary that will be utilized to represent the relation between Subject and Object, which in turn constitutes the Predicate.

To index the lines, we present a table with the lists of columns of the data conjunction. The two first columns are used to identify the field of the data conjunction and to enable the indexing of data in those fields. It is not mandatory to index all the fields – the Publication Agent must decide, according to its knowledge about the data conjunction, whether to index the field or not, by deciding that either the data is relevant to the indexing or that it is redundant from the data connection.

On the next column of the table, the Predicate that describes the relation between Subject and Object is chosen. In that moment, the Publication Agent selects which vocabulary to use to describe the relation coming from the terms existing on the metadatas and/or ontologies selected in the previous step.

The fourth column of the table is the definition of the triple object. The first option to chose from is that the object is Literal. With that option, raw data coming from the conjunction of data constitutes the Object. If the Object is not Literal, the Publication Agent will have the option to select the RDF file registered on the UnBGOLD that possesses the same vocabulary defined in the previous step. By selecting the RDF file it presents, to the side, the options for the data conjunction fields. This is precisely the moment when users can connect a data with the other conjunctions of data, because the tool will search inside the RDF file for the resource that will be interacting with that data. If there is no resource, the subject is generated as Literal.

In the example presented, the indexing table has three lines that represent the columns from the dataset. The first line is the "code" field , which was not chosen to be indexed, since this option disables the fields utilized referring to the line in question on the table. In the second line, we have the "name/major" field. In this case, the "dc:Title" vocabulary was selected as a Predicate and we chose Literal for the Object. In the last line, the "department" field was used by the "aiiso:Departament" Predicate, since the Subject of this triple was chosen as another resource. In this case, the "departament_3.rdf" file was selected, where it was enabled a field by its side where the Publication Agent selected the "name" field. This means that the tool connects the Subject of that triple through an existent resource inside the RDF file selected where the value of the "name" field of the file has to be the same as the value from the "department" field in the dataset presented.

Figure 10 presents the result of an indexed line from the example presented. An XML code can be observed and the connection graph

[11] http://dados.gov.br/pagina/cartilha-publicacao-dados-abertos

Figure 9: Semantic Indexing Step

in which one can check which IRI identifies the triple's resource, the vocabularies utilized and the Objects, given that there is the "College of Administration" Literal and also the "http://adm.unb.br" resource as Object.

The generating of the RDF file from the request to programmed publication or by manual solicitation of the Publication Agent that will be able to make the file's download.

```
1: <?xml version="1.0"?>
2: <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3:   xmlns:dc="http://purl.org/dc/elements/1.1/"
4:   xmlns:aiiso="http://purl.org/vocab/aiiso/schema/">
5:   <rdf:Description rdf:about="http://unb.br/cursos/10">
6:     <rdf:type>aiiso:Course</rdf:type>
7:     <dc:title>Faculdade de Administração</dc:title>
8:     <aiiso:Departament>http://adm.unb.br</aiiso:Departament>
9:     <sigla>ADM</sigla>
10:  </rdf:Description>
11: </rdf:RDF>
12:
```

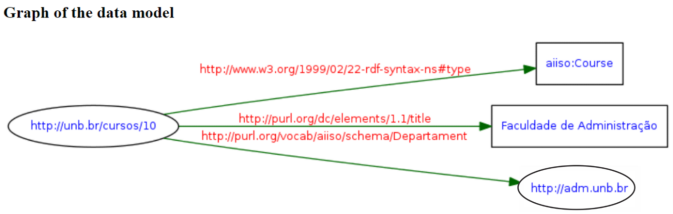

Figure 10: Triple Course and Department

4 CONCLUSIONS

Before implementing open data policies in Brazil, it is important to make processing this data viable and as easy as possible. Thus, we propose such a tool that also adds quality through the semantic indexing of this data to guarantee a relation of this data making the open data connected. Today the majority of datasets available are published in isolated form where the integration of different bases is harder because of a lack of standardization. The UnBGOLD is a tool to aid institutions that want to index their data through a friendly interface.

The UnBGOLD was proposed inside the architecture of open data publishing of the UnB, being that this architecture has the characteristic of being uncoupled, where the log in and log off interfaces can be replaced by the architecture of other organs that start using this tool without having to perform huge adjustments in the already existent process, aside from making the automation of the publishing possible, if that is the institution's interest.

Some of the advantages of using this tool, include a decrease in human intervention in the publishing process, which reduces the possibility of error and making it possible for the data to be published at shorter intervals, and as recent as possible, and at the same time increasing publishing quality with the data semantic increment. However, it is important that the initial curation of the first publication of a data set, if it serves as a basis for automation, is carried out considering the data updating requirements.

In this context, publishing connected open data is facilitated in such a way that with the evolution, the connection between different datasets becomes the natural path of the publications. Thus, this data is recognized as being the highest quality, achieving 5 stars in the index proposed by Bernie Lee. With the publication of the connected open data, the possibility of semantic enrichment in data searches will bring more quality to the results, improving analyses and increasing the transparency of public management.

REFERENCES

[1] [n. d.]. Padrões de metadados.
[2] Everton Agilar, Rodrigo Almeida, and Edna Canedo. 2016. A Systematic Mapping Study on Legacy System Modernization. In *SEKE*.
[3] Tim Berners-Lee. 2006. Desing Issues - Linked Data. https://www.w3.org/DesignIssues/LinkedData.html
[4] Tim Berners-Lee, James Hendler, Ora Lassila, et al. 2001. The semantic web. *Scientific american* 284, 5 (2001), 28–37.
[5] Christian Bizer and Richard Cyganiak. 2014. RDF 1.1 TriG-RDF Dataset Language-W3C. (2014). https://www.w3.org/RDF/
[6] Brasil. 2011. LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm
[7] Brasil. 2016. DECRETO Nº 8.638 DE 15, DE JANEIRO DE 2016. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm
[8] David Eaves. 2009. The Three Laws of Open Government Data. http://eaves.ca/2009/09/30/three-law-of-open-government-data/
[9] Apache Software Foundation. 2013. Getting started with Apache Jena. http://jena.apache.org/getting_started/index.html
[10] OKF Open Knowledge Fundation. 2015. The Open Definition.
[11] Durval Vieira Pereira. 2017. Modelagem e representação semântica de dados governamentais abertos da previdência social brasileira. (2017).
[12] Sebastopol Group. 2007. 8 Principles of Open Government Data. http://public.resource.org/8_principles.html