19a21novembro2014

Florianópolis, SC

ANAIS

LOD BRASIL
LINKED OPEN DATA

ISBN 978-85-61115-09-8

UFSC

EGC

INSTITUTO FEDERAL
SANTA CATARINA

ISBN 978-85-61115-09-8

Fernando Álvaro Ostuni Gauthier

Antônio Pereira Cândido

José Leomar Todesco

Marco Antonio Neiva Koslosky

Silvia Maria Puentes Bentancourt

Cleverson Tabajara Vianna

Organizadores

LOD BRASIL

Linked Open Data

Anais

Florianópolis

UFSC/EGC

2014

ISBN 978-85-61115-09-8

Edição e Diagramação

Cleverson Tabajara Vianna

Silvia Maria Puentes Bentancourt

Capa

Geisa Golin Albano

Catalogação na fonte pela Biblioteca Universitária

da Universidade Federal de Santa Catarina

C7491 Congresso Linked Open Data Brasil (1. : 2014 : Florianópolis, SC).
LOD Brasil : Linked Open Data : Anais [recurso eletrônico]/ I Congresso Linked Open Data Brasil; organizadores, Fernando Álvaro Ostuni Gauthier...[et al.]. - Florianópolis : UFSC/EGC, 2014.
298 p., il., graf., tabs.

Modo de acesso:<http://lodbrasil.com.br/>
Inclui bibliografia.
Evento realizado de 19 a 21 de novembro de 2014.
ISBN 978-85-61115-09-8

1. Tecnologia da informação. 2. Web semântica 3. Gestão do conhecimento. I. Gauthier, Fernando Álvaro Ostuni. II. Título.

CDU: 004

EGC

NDC

ISBN 978-85-61115-09-8

Organização

Coordenação Geral
Fernando Álvaro Ostuni Gauthier, Dr.
Antônio Pereira Cândido, Dr.
Secretaria: Evelin Trindade

Comitê Operacional
Coordenador: Cleverson Tabajara Vianna, Me.
Alexandre Gonçalves, Dr.
Rafael Speroni, Me.
Adriano Heis, Me.

Comitê Financeiro
Fernando Álvaro Ostuni Gauthier, Dr.
Marco Antonio Neiva Koslosky, Dr.
Marcelo Macedo, Dr.

Comitê Comunicação
Coordenação: Clarissa Stefani e Waléria Kulkamp Haeming, Dr.
Maricel Torres
Evelin Trindade
Daniel Fernando Anderle
José Alcino Furtado
Geisa Golin Albano
Giovana Perine Jacques

Comitê Científico
Coordenador: José Leomar Todesco, Dr. – Universidade Federal de Santa Catarina (UFSC)
Sub-Coordenador: Marco Antonio Neiva Koslosky, Dr. (IFSC)
Prof. Dr. Aires Rover – Universidade Federal de Santa Catarina (Brasil)
Prof. Dr. Alexandre Leopoldo Goncalves – Universidade Federal de Santa Catarina (Brasil)
Prof. Dra. Ana Maria de Carvalho Moura – Laboratório Nacional de Computação Científica (Brasil)
Prof. Dr. Antônio Pereira Cândido – Instituto Federal de Santa Catarina (Brasil)
Prof. Dra. Asuncion Gomez-Perez – Universidade Politécnica de Madri (Espanha)
Prof. Dra. Bernadette Farias Loscio – Universidade Federal de Pernambuco (Brasil)
Dr. Boris Villazon-Terrazas – iSOCO, Intelligent Software Components (Espanha)
Prof. Dr. Denilson Sell – Universidade do Estado de Santa Catarina (Brasil)
Prof. Dr. Edgard Marx – University of Leipzig (Alemanha)
Prof. Dr. Fernando Alvaro Ostuni Gauthier – Universidade Federal de Santa Catarina (Brasil)
Prof. Dr. Fernando Parreiras – Universidade FUMEC (Brasil)

ISBN 978-85-61115-09-8

Prof. Dr. Jean-Paul Calbimonte – École polytechnique fédérale de Lausanne (Suiça)
Prof. Dr. Jose Francisco Salm Jr. – Universidade do Estado de Santa Catarina (Brasil)
Prof. Dr. Jose Leomar Todesco – Universidade Federal de Santa Catarina (Brasil)
Prof. Dr. Jose Maria Parente de Oliveira – Instituto Tecnológico de Aeronáutica (Brasil)
Prof. Dr. Juliano Brignoli - Instituto Federal Catarinense (Brasil)
Prof. Dr. Leandro Krug Wives – Universidade Federal do Rio Grande do Sul (Brasil)
Prof. Dra. Lia Caetano Bastos – Universidade Federal de Santa Catarina (Brasil)
Prof. Dr. Luiz Antônio Moro Palazzo – Universidade Católica de Pelotas (Brasil)
Prof. Dr. Marcello Peixoto Bax – Universidade Federal de Minas Gerais (Brasil)
Prof. Dr. Marco Antonio Neiva Koslosky – Instituto Federal de Santa Catarina (Brasil)
Prof. Dr. Marcio Vieira de Souza – Universidade Federal de Santa Catarina (Brasil)
Prof. Dr. Mario Antônio Ribeiro Dantas – Universidade Federal de Santa Catarina (Brasil)
Prof. Dr. Nelson Piedra – Universidad Técnica Particular de Loja (Equador)
Dra. Nuria García-Santa – iSOCO, Intelligent Software Components (Espanha)
Prof. Dr. Oscar Corcho Garcia – Universidade Politécnica de Madri (Espanha)
Prof. Dr. Renato Fileto – Universidade Federal de Santa Catarina (Brasil)
Prof. Dr. Roberto Carlos dos Santos Pacheco – Universidade Federal de Santa Catarina (Brasil)
Prof. Dr. Rogério Cid Bastos – Universidade Federal de Santa Catarina (Brasil)
Prof. Dr. Sandro Rautenberg – Universidade Estadual do Centro-Oeste (Brasil)
Prof. Dra. Vania Ulbricht – Universidade Federal de Santa Catarina (Brasil)
Prof. Dr. Victor Saquicela Galarza – Universidad de Cuenca (Equador)

UNIVERSIDADE FEDERAL DE SANTA CATARINA

INSTITUTO FEDERAL SANTA CATARINA

EGC

LatinAmerica Linked Data

ISBN 978-85-61115-09-8

Apoio

FAPESC

FEPESE

Apresentação

O **Congresso *Linked Open Data* Brasil** é uma iniciativa pioneira do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (EGC) que busca congregar pesquisadores da área no Brasil.

Os temas do congresso abrangem questões teóricas e aplicações de dados abertos e dados ligados em organizações públicas e privadas.

Os anais apresentam os trabalhos selecionados no sistema peer review para apresentações orais e pôsters.

Com este evento, o EGC espera contribuir para a consolidação da área e de redes de colaboração entre pesquisadores e instituições.

A difusão da cultura de dados abertos e ligados na sociedade brasileira é outro dos objetivos do evento.

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

Sumário

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

IMPROVING SEARCH RESULTS IN THE LOD WITH ENTITY RESOLUTION

Gustavo de Assis Costa
gacosta@gmail.com

José Maria Parente de Oliveira
parente@ita.br

Resumo

A capacidade de expressividade de consultaspode ser alcançada, dentre as várias características existentes,principalmente devido à existência de ligações entre diferentes bases de dados. Neste sentido, a capacidade de agregação de dados pode ser um diferencial para alguns motores de busca em dados ligados. No entanto, existem muitos desafios, especialmente quando se considera diferentes tipos, estruturas e vocabulários utilizados na Web. Além disso, é difícil garantir a qualidade dos dados, porque eles são geralmente incompletos, inconsistentes e contêm valores discrepantes. Tentando superar alguns desses problemas, muitos trabalhos têm aplicado a tarefa de Resolução de Entidades utilizando diferentes técnicas e algoritmos. Neste artigo apresentamos uma visão geral da nossa experiência obtida na construção de uma abordagem para integrar diferentes bases de dados com o objetivo de melhorar os resultados de pesquisa feita sobre a nuvem LOD. Além de uma descrição geral da abordagem e seus aspectos principais, apresentamos alguns resultados preliminares obtidos, e um breve panorama sobre ER e algumas tendências sobre o potencial desta técnica.

Palavras-chave: Artigo Científico. Metodologia. Normas.

Abstract

Expressive query capabilities can be achieved, among many other features,mainly due to the traversal of links between different data sources. In this sense, the ability of data aggregation could be a differential to some Linked Data search engines when crawling the Web of Data. However, there are many challenges specially when considering different types, structures and vocabularies used in the Web. Besides that, it is difficult to guarantee the quality of data because they are usually incomplete, inconsistent and contain outliers. Trying to overcome some of these problems, many works have applied the task of Entity Resolution using different techniques and algorithms. In this paper we present an overview of our experience obtained in the construction of an approach to integrate data sets aiming to improve search results made over the LOD. In addition to a general description of the approach and its main features, we present some preliminary results, and a brief overview of ER and some trends about the potential of application of this technique.

Key Words: Linked Data, Semantic Web, Entity Resolution

ISBN 978-85-61115-09-8

Introduction

Following the trend of the World Wide Web, a considerable number of people and companies have published their data in the Web of Data (HEATH AND BIZER, 2011). As a result, the amount and variety of data is growing exponentially, creating a graph of global dimensions formed by billions of RDF triples that represent data from different fields of knowledge.

The creation of RDF data sources are commonly based in conversion processes from structured data that comes from relational databases or from semi-structured or unstructured data crawled from web pages, texts and other type of documents. As these data sources typically present problems like outliers, duplication, inconsistency, and other, like schema heterogeneity, derived data will have them in the same way. Among the existing problems, these arise as some of the limiting factors to the effective integration and sharing of Linked Data.

To deal with these problems, methods of inductive approaches from the fields of machine learning and data mining were successfully employed to perform approximate reasoning and to derive predictions which are neither explicitly asserted in the knowledge base nor provable based in logical reasoning (RETTINGER et. al. 2012; TRESP et. al. 2008). In general, some tasks can be performed: classification (object type prediction or property value prediction), link prediction, clustering and ER. In addition to the previously mentioned approaches, the Semantic Web community recognizesthe approach of instance-level ER(BIZER, HEATH, and BERNERS-LEE, 2009). In this way, methods often make use of similarity metricsapplied between entities based on established techniques from database community, like record linkage or de-duplication (FELLEGI and SUNTER 1969), and from ontology community, like ontology matching (EUZENAT and SHVAIKO 2007).

Following research trends found in the literature, most of works concentrated in methods and techniques related to the discovery of links between entities in the Web of Data in the following ways: i) the discovery of resources from different datasets that represent the same real world object, a.k.a. data linking (ER and ontology matching (FELLEGI and SUNTER, 1969; ELMAGARMID, IPEIROTIS, and VERYKIOS, 2007; EUZENAT and SHVAIKO 2007; JAIN et. al., 2010; JEAN-MARY, SHIRONOSHITA, and KABUKA, 2009; WINKLER, 2006)) and ii) predicting the probability of RDF links existence based on implicit patterns found in the data (statistical relational learning (GETOOR, 2003; GETOOR and DIEHL, 2005; RETTINGER et. al., 2012; TRESP et. al., 2008)).

From this viewpoint, many works concentrated in try to overcome problems by addressing the task of ER which deals with extracting, matching and resolving entity mentions in structured and unstructured data. In linked data research community it is recognized as a prominent issue. Also known as record linkage, de-duplication, co-reference resolution, instance matching, among others, it has been used to look for interrelationships, previously unknown, between different representations of the same real world entity.

From this scenario, some posed challenges deserve attention. The first is deal with semi-structured data. Different semantic description structures can be employed to refer to the same element, as for example, the description of entities of the same type. As an example, in addition to naturally different namespaces descriptions, each vocabulary can employ different descriptions to refer to the same structural type. This reflects a classical problem of ontology alignment (EUZENAT and SHVAIKO, 2007), which aims to look for mappings between different schemas.

A second challenge is related to noise in the data. As already stated, there are various problems related to literal descriptions of data. To overcome these problems, many different metrics of string matching were proposed. Even taking into account some situations like

ISBN 978-85-61115-09-8

characters suppression, variation in radical words, among others, existent metrics still cannot resolve problems like attributes without value.

Finally, the third challenge is deal with scale. The LOD cloud has nowadays about 60 billion RDF triples. Handle this amount of information requires techniques that can parallelize the process, as for example Map Reduce. However, there are still few works that discusses the application of these techniques in problems of ER.

In this paper we aim to report our experience in developing an approach based in a statistical approximation method that captures joint evidence of similarity related to values of descriptions of entities and any relationship correlation through existing entity-entity predicates. This way, our main contribution is to provide subsidies so that researchers can have a starting point in developing new approaches and applications that may use ER.

The proposed approach focuses on considering most of problems presented above and it is divided in four steps: 1) Preprocessing, 2) Pair-wise string similarities 3) Similarity evidences modeled as a matrix entity-attribute 4) Relationships between entities modeled as a tensor. Finally, the problem is formulated as a coupled matrix and tensor factorization. The work was supported by applying an extended version of RESCAL (NICKEL, TRESP and KRIEGEL, 2011 (a)) model, a tensor factorization model for relational learning.

The remainder of this paper is structured as follows: Section 1 presents an overview of the approach, while in Section 2 we summarize related works. Section 3 describes an overview about the solution space of ER and a framework of related concepts with the topic. In the last section we make some discussions and conclude with some notes on future research trends.

1 The relational approach

Figure 1 depicts an overview of our approach. The literal information for each entity in all datasets are extracted. There are some discriminative literal descriptions that, considering the triple structure, extracts the most significant information of each element:

- Attribute values. Correspond to features of an entity (e.g. name/label, birth date, profession). Most approaches explore these values due to the precision when identifying an entity;
- URI infix. In (PAPADAKIS et. al., 2010) experiments results showed that approximately 66% of the 182 million URIs of a dataset follow a common pattern: the Prefix-Infix(-Suffix) scheme. Each component of this form plays a special role: the Prefix part contains information about the source (i.e., domain) of the URI, the Infix part is a sort of local identifier, and the optional Suffix part contains either details about the format (e.g., .rdf and .n3), or a named anchor.
- Predicate: We will use the last token (normalized) of the URI, e.g., “has spouse” for "fb:has_spouse".

Considering two datasets A and B with n and m entities respectively, and a brute force algorithm, there will be at least n x m comparisons between instance pairs. It is impractical, especially when dealing with real world datasets, with millions of triples or even larger. To overcome this problem, we perform a preprocessing step to obtain the possible matching pairs. An inverted index is built for instances of some key words in the descriptions to efficiently determine potential candidates. The entities sharing the same keys in the index are considered to be candidate matching instances.

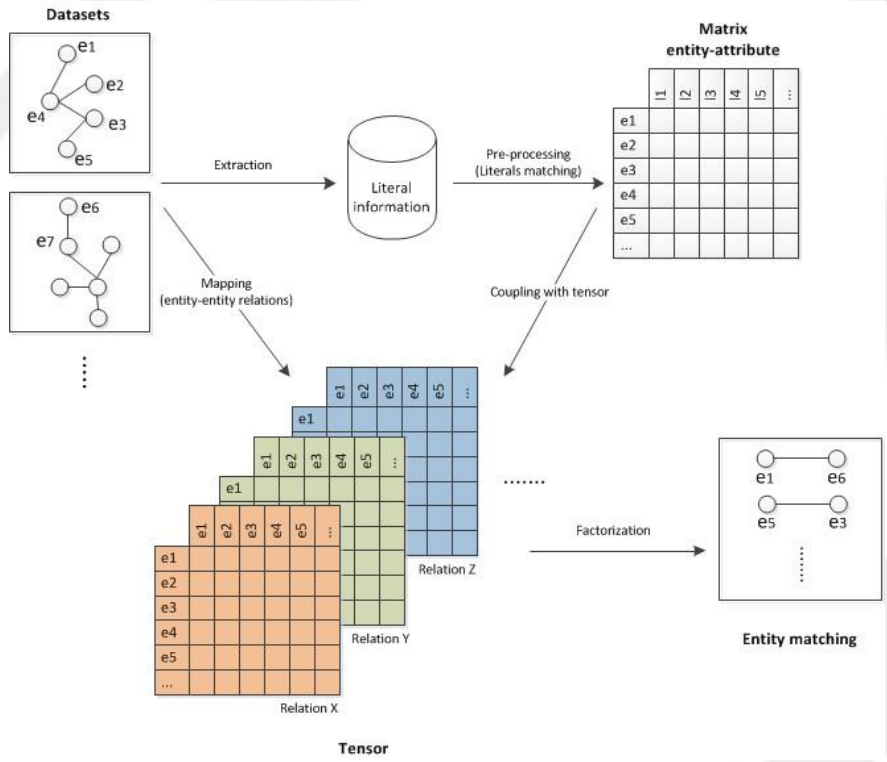
After literal information extraction, we apply the similarity metrics for the candidates and generate the entity-attribute matrix, with its entries corresponding to an entity having or not certain attribute or an infix in its URI description. In the same way, entity-entity relations

ISBN 978-85-61115-09-8

from datasets are mapped to tensor, with which we perform the coupled factorization. The factor-matrix A computed in the above process can be interpreted as an embedding of the entities into a latent-component space that reflects their similarity over all relations in the domain of discourse.

In order to retrieve entities that are similar to a particular entity e with respect to all relations in the data, we compute a clustering in the latent-component space. Initially, however, we normalize the rows of A, such that each row represents the normalized participation of the corresponding entity in the latent components. From feature vectors corresponding to each entity (matrix rows) it is possible to create clusters of similar entities, since matrix A represents entities by their participation in the latent components. The clustering will be determined by the entities' similarity evidences in the relational domain.

Fig. 1. An overview of the proposed framework



1.1 Information Extraction

We try to extract the greatest number of literal information from the RDF triples. The literals can be grouped in three types: a) Attribute values; b) URI Infixes and c) Predicates. Attribute values can be a feature, a description or even an associated event, that when considered jointly, can uniquely identify an entity. It's important to point out that not always we have all the values informed or even they can be inconsistent.

A second type of literal that can be of relevance to the process are the URI infixes of subjects or objects within triples. As stated in(PAPADAKIS et. al., 2010), it could be expected that Infixes of URIs, which are more source-independent than the Prefixes and the Suffixes, can contain the most discriminative information for the similarity task within a URI. Despite the high heterogeneity in the Prefixes of the URIs, the Infix remains the same. The Suffix is optional and can be ignored when matching URIs. Table 1 illustrates an example with two URIs that refer to the same person but are syntactically different. Specifically in this

ISBN 978-85-61115-09-8

case, even the infixes are different, necessarily requiring the application of a string similarity metric.

Table 1. Example of two URIs that refer to the same person

Prefix	Infix	Suffix
http://liris.cnrs.fr	/olivier.aubert	/foaf.rdf#me
http://bat710.univ-lyon1.fr	/~oaubert	/foaf.rdf#me

Fonte: PAPADAKIS et. al., 2010

By definition, a predicate is the second part of an RDF statement and defines the property for the subject of the statement. Unlike a subject or object, a predicate must always be a URI. From an empirical analysis it appears that property matching is not trivial since the datasets were usually designed with their own ontologies. Nevertheless, if we make analyzes of synonymy it is possible to overcome results obtained only with similarity metrics. Some other linguistic facts, like polysemy and homonymy, are not treated because the isolated weight of these phenomena have little significance when considering all the contextual facts involved in the likelihood estimation of similarity in the model.

1.2 Strategies for evaluating and modeling the similarity

Some different similarity metric functions can be used for different types of literal information. Each of the literals have its own characteristics and thus we consider the application of different metrics and strategies to analyze string similarity.

For attribute values we use TF-IDF. An attribute can contain a great diversity of values, with different sizes and types. When performing the algorithm we will treat each string/label relative to entities as a set of values (documents). The TF-IDF weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.

As URI infixes are normally composed of little strings to represent a single identifier, we decided to use Edit-Distance metric. It is a metric that measures the distance between two words as the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other.

Finally, it's important to highlight that predicates frequently involve verbs, which can appear in a wider variety of forms than nouns. They also contain often more functional words, such as articles and prepositions. In this case we try to overcome the difficulty by using a combination of the metric of Edit-distance with the Wordnet to perform a jointly analysis that considers besides the string similarity, an analysis of synonyms between the two descriptions. In this strategy, when the similarity metric result is below to some predefined threshold, we apply the analysis of synonymy.

After processing all the similarities we have to model the evidences in the entity-attribute matrix. The matrix D of size n x l is composed of n entities at rows and l literal evidences at columns. A matrix entry $D_{ij} = 1$ denotes that an entity have certain attribute value. Otherwise, if the entity does not have this attribute it will be set to 0. If two entities share the same attribute, but with different attribute values, they will not share the same entry in the table, i.e, each of them will have its distinct entry set to 1. In the sense of normalizing a set of similar values according to the metric, each column will be represented by one canonical value randomly chosen from the set, i.e., whether the values are slightly different, there will be just one entry . Although URI infixes and predicates are not strictly attribute values, we will handle them the same way as attributes. As a result, the matrix will contain all

ISBN 978-85-61115-09-8

the literal evidences obtained from the performance of different metric and strategies of similarity.

1.3 Statistical relational learning model

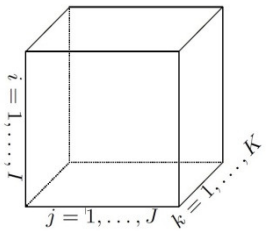
We now present the model that we applied in our approach. Firstly, we need to identify the key elements of the model which in turn are the entities and its relations. The entities are given by the set of all resources, classes and blank nodes in the data, while the set of relations consists of all predicates that include relationships between entities. Once these elements were extracted from datasets we set out to the transformation of them into a tensor representation.

A tensor is a multidimensional array. More formally, an N-way or Nth-order tensor is an element of the tensor product of N vector spaces, each of which has its own coordinate system. A third-order tensor has three indices as shown in Figure 3. A first-order tensor is a vector, a second-order tensor is a matrix, and tensors of order three or higher are called higher-order tensors (KOLDA and BADER, 2009).

Assuming that our relational domain consists of n entities and m relation types, data is modeled as a three-way tensor X of size$n \times n \times m$,, where the entries on two modes (dimensions) of the tensor correspond to the combined entities of the domain of discourse and the third mode holds the ^{m} different types of relations.

A tensor entry X_{ijk}= 1 denotes the fact that the k-th relation (i-th entity, j-th entity) exists. Otherwise, for non-existing or unknown relations, X_{ijk} is set to zero.

Fig. 2. Tensor model for relational data. i and j represent entities and k the relationships.



Besides the modeling aspect, the motivation to use tensor factorization is due to its power of prediction when used as machine learning task, as is done in SVD method for example. The process of factorization decomposes an observed matrix/tensor into latent (or hidden) factors. Latent factors can be interpreted as new features that have been invented to describe the data.

In RESCAL, learning is performed using the latent components of the model (Fig. 3a). The approach employs the rank-r factorization as follows, where each segment is factored as X_k

$$X_k \approx AR_kA^T \text{ where } k = 1, \ldots, m \quad (1)$$

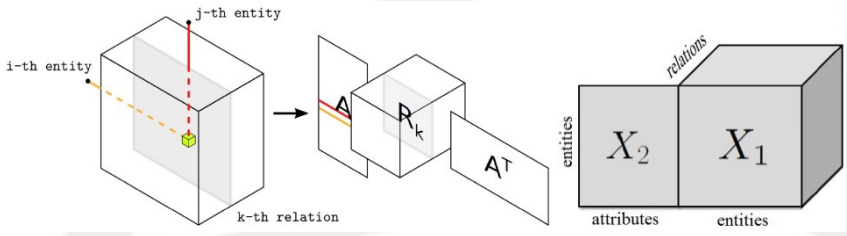
A is a $n \times r$ matrix containing the components of the latent representation of the entities in the domain and R_k is an asymmetric matrix $r \times r$ modeling the interactions of the components of the k_{th} latent predicate. The rows of the factor matrices A and R can be considered latent-variable representations of entities that explain the observed variables X_{ij}, the columns can be considered the invented latent features and the entries of the factor matrices specify how much an entity participates in a latent feature.

The factor-matrices A and R_k are computed by solving a regularized minimization problem (NICKEL, TRESP and KRIEGEL, 2011 (a))applying an alternating least squares

ISBN 978-85-61115-09-8

algorithm (RESCAL-ALS), which updates A and R_k iteratively until a convergence criterion is met (linear regression). In order to retrieve entities that are similar to a particular entity e with respect to all relations in the data, it is sufficient to compute a ranking of entities by their similarity to e in A.

Fig. 3. a) Illustration of data representation and factorization in the model and b) A Tensor coupled with a matrix of attributes



Fonte: NICKEL, TRESP and KRIEGEL, 2011 (a)

Once this model assumes that two of the three modes are defined by entities, the process becomes limited to RDF resources. So we used an extension of the model, coupling the entity-attribute matrix with the tensor (Fig. 4b) aiming to perform the factorization (NICKEL et al., 2012; YILMAZ et al., 2011).

If we include all the literal evidences in the tensor, a huge amount of entries would be wasted, which would lead to an increased runtime since a significantly larger tensor would have to be factorized. So, the idea is to add the predicate-value pairs to a separate entity-attributes matrix D and not to the tensor X. The entity-attributes matrix D is then factorized into

$$D \approx AV \tag{2}$$

where A is the entities' latent-component representation of the model and $V$is an $r \times l$matrix, which provides a latent-component representation of the literals. To include this matrix factorization as an additional constraint on Ain the tensor factorization of X, it is necessary to adjust the minimization problem.

In figure 4 we show an illustration that depict an example. The latent-component representations of entities A and B will be similar to each other in this example, as both representations reflect that their corresponding entities are related to the same object (wikipedia page) and attribute value. Because of this and their own similarity evidences, C and D will also have similar latent-component between their representations. Consequently, the latent feature vector of A will yield similar values to the latent feature vector of B and as such the likelihood of matching can be predicted correctly. The attribute values are only considered here due to the extension of the model.

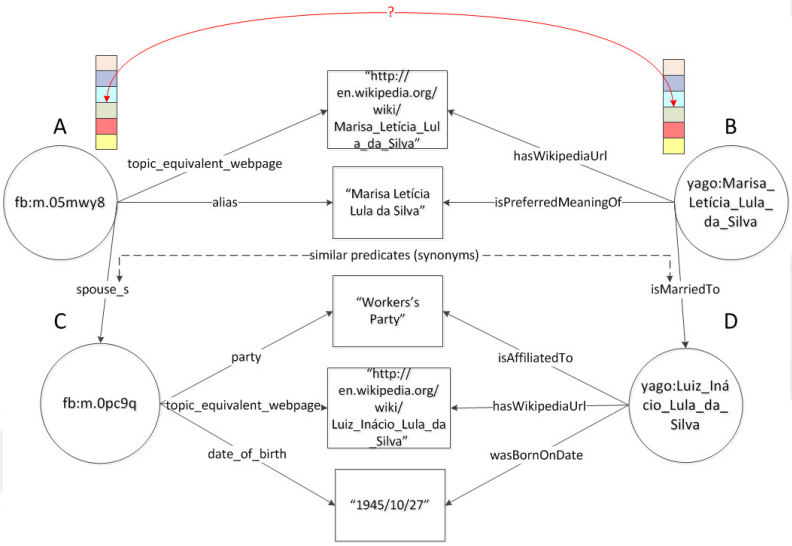
Considering that a_i and a_j denote the i-th and j-th row of A and thus are the latent-component representations of the i-th and j-th entity, the products

1) $a_{\text{fb:m.05mwy8}}{}^{T}R_{\{\text{spouse\_s, isMarriedTo}\}}\ a_{\text{fb:m.0pc9q}}$,
2) $a_{\text{fb:m.05mwy8}}{}^{T}R_{\{\text{spouse\_s, isMarriedTo}\}}\ a_{\text{yago:Luiz\_Inácio\_Lula\_da\_Silva}}$,
3) $a_{\text{yago:Marisa\_Leticia\_Lula\_da\_Silva}}{}^{T}R_{\{\text{spouse\_s, isMarriedTo}\}}\ a_{\text{fb:m.0pc9q}}$
4) $a_{\text{yago:Marisa\_Leticia\_Lula\_da\_Silva}}{}^{T}R_{\{\text{spouse\_s, isMarriedTo}\}}\ a_{\text{yago:Luiz\_Inácio\_Lula\_da\_Silva}}$

along with the similarities evidences obtained, will contribute to get likelihood of A and B representing the same real world entity.

ISBN 978-85-61115-09-8

Fig. 4. Illustration with representations of the same real world entities in Freebase and YAGO. The red line indicates the wanted matching.



2 Related Work

As stated before, the problem of ER has been exploited in different research fields that is why there are many different approaches.Some approaches are based on domain specific solutions. One first tool is SILK - Link Discovery Framework (VOLZ et. al., 2009). It is a well-known tool for publishing and managing RDF relationship between two RDF datasets. The framework provides a declarative language in which different string similarity metrics, defined by the user, can be manually combined. Raimond et. al.(RAIMOND, SUTTON and SANDLER, 2008) addresses the problem in the domain of music, modeling datasets as graphs, performing mapping between the graphs. Sleeman and Finin (SLEEMAN and FININ, 2010) and Shi et. al.(SHI et. al., 2008) present solutions for solving FOAF entities using logical constraints. Another group of solutions are domain independent.

Among them, there are some papers that addresses logical constraints like functional/inverse functional properties and cardinality as key aspects in their solutions (IOANNOU et. al., 2010; HOGAN et. al., 2012; HU, QU and SUN, 2011; NIU et. al., 2011). Using this properties is not sufficient to find traces of similarity in LOD. Hogan et. al. (HOGAN et. al., 2010) tries to find more inverse functional properties with a statistical method. However, datasets must share the same vocabulary.

Some papers focus on improving the efficiency on the matching. Ngomo and Auer (NGOMO and AUER, 2011) present LIMES framework in order to circumvent the scalability problem by applying the method of triangulation in metric spaces. Song and Heflin (SONG and HEFLIN, 2011) generate candidates by indexing some key words of instances. To our knowledge, as our work, few papers (PAPADAKIS et. al., 2010; PAPADAKIS et. al., 2013; BÖHM et. al., 2012) focus on improving the effectiveness of matching with RDF data.

Just recently some few approaches have started to use the parallelization strategy for ER. Some works addresses the blocking technique using a sorted neighborhood approach (KOLB, THOR and RAHM, 2012b) and in (KOLB, THOR and RAHM, 2012a) it is addressed as a load-balanced ER. The basis of these works is a map function that emits <key, value> pairs and pairs with the same key are processed in the same reducer, and in the reduce function the matching function is performed with all pairs that have the same key.Like most

ISBN 978-85-61115-09-8

of the existing works in the scientific literature, none of these approaches addresses graph data, like RDF. RDF is essentially semi-structured data and it is naturally dirty, incomplete and inconsistent. Despite the work in (PAPADAKIS et. al., 2014; PAPADAKIS et. al., 2012)take into account the intrinsic issues of semantic web data, they do not address parallelization techniques like MapReduce in its approaches.

Linda system (BÖHM et. al., 2012) is an approach that addresses RDF data and its algorithm is based on maintaining X and Y matrices as data structures. Matrix X will contain matching or non-matching results from performed comparisons that are temporarily maintained in matrix Y, and this second contains real-valued similarity values. The algorithm then repeatedly dequeues entity pairs with the highest similarity score from a priority queue. It considers the nearest neighborhood of matching entities as a way to propagate similarities between linked entities.

3 An overview of the solution space

The solution space for the ER problem can be quite large due to the many different possible treatments in all the contexts. This way, it is necessary consolidate certain information and then obtain an easier way to address the problem.

To facilitate understanding of the problem we created a conceptual map (Fig. 6) that depicts our vision about it. We tried to outline the specificities of ER task in the context of RDF in Linked Data.In this sense, first of all we can mention three types of input to the method: tabular, tree and graph data. These classification is based in the different degrees of structuredness of each type of data which directly influence the difficulty of the methods.

- Tabular: data with high structuredness. Compare values of the same attributes is enough to compute similarities. Ex.: relational databases;
- Tree data: data with varying structuredness. Similarity of values is affected by the similarity of their ancestors and descendants. Ex.: XML;
- Graph data: data with varying structuredness. Computing similarities becomes harder. Ex.: RDF data, which present cycles and non-unique root elements.

With regard to the type of method, there are three different ways of addressing the problem using different techniques and algorithms:

- Iterative methods (KIM and Lee 2010): identify matches that can lead to new matches, using, for example, the already merged descriptions. In this method there will be more matches than in the others;
- Blocking methods (PAPADAKIS et. al., 2012; PAPADAKIS et. al., 2013): group together descriptions close to each other. These methods rely on criteria for placing descriptions into blocks (blocking keys) aiming to reduce the number of comparisons over the cost of missing matches;
- Learning methods (RONG et. al., 2012; NGUYEN, ICHISE and LE, 2012; COSTA and PARENTE DE OLIVEIRA, 2014(a); COSTA and PARENTE DE OLIVEIRA, 2014(b)): use of training data, annotated as matches or not. Classify descriptions, using statistical inference.

Finally, all the methods have at least one objective which is essential in the choice of which method to use.

- Effectiveness: find as many (few) true (false) matches as possible. It is achieved by increasing the number of comparisons in a single or multiple iterations. Targeted by iterative methods.
- Efficiency: resolve the given descriptions as fast as possible, e.g. by reducing redundant comparisons. Pre-processing to place descriptions in blocks. Targeted by blocking methods.

ISBN 978-85-61115-09-8

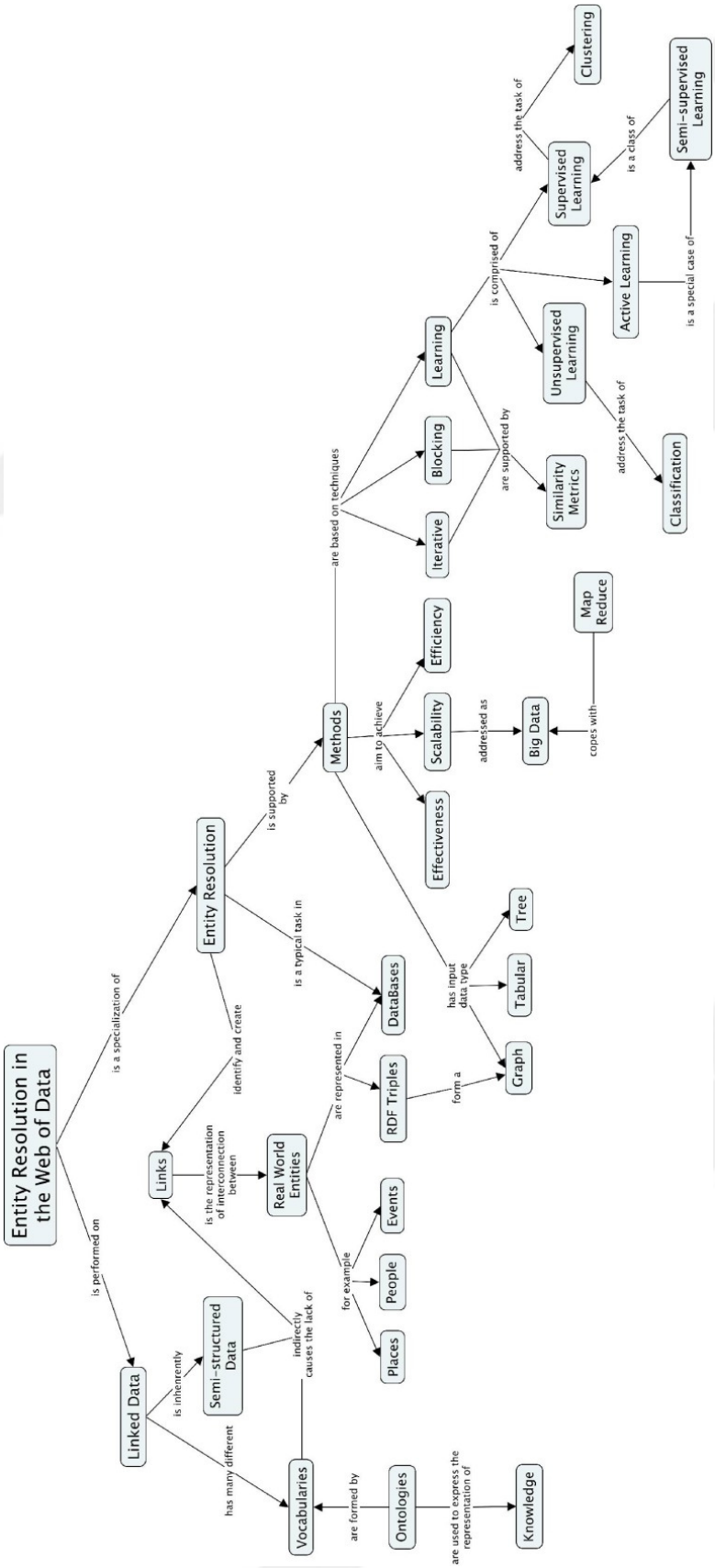
– Scalability (COSTA and PARENTE DE OLIVEIRA, 2014(b)): Methods that can cope with Big Data. Distribute the task of ER to multiple computational resources, e.g. Map-Reduce

In our work we focus in an hybrid method type that can explore the efficiency of blocking methods jointly with the effectiveness usually addressed in iterative methods. When we used metrics to populate the entity-attribute matrix we applied the inverted index technique to avoid comparisons between non-potential matchings, as is done with blocking methods. The effectiveness is considered when we model data in a tensor with all the literal evidences jointly with all the relations between entities. This way, we explore all existent information in datasets.

Considered as a learning method, our approach is different from all other because it is not supported by classification task. Unlike this, it makes use of an algorithm that is based in statistical relational learning, performing inferences using statistical approximation with all latent variables of the model. The greatest advantage of it is its ability to consider entity relationships.

ISBN 978-85-61115-09-8

Fig. 5 -A conceptual map of ER in Web of Data



ISBN 978-85-61115-09-8

Results and Discussion

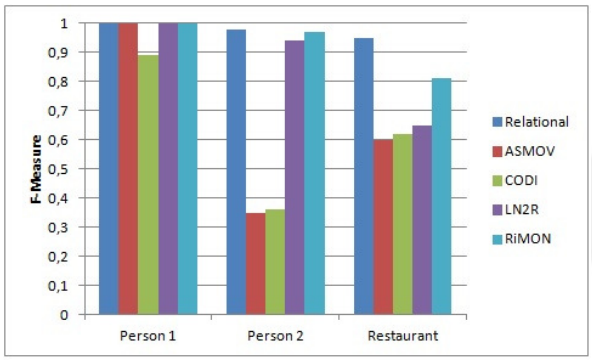
Aiming to get the most out of the different data sets, it is necessary to provide integration of all of them. In this sense, ER is one of the most prominent method when we consider improvement of search results, mainly because of the maturity of scientific community. Furthermore, there are so few available resources that allow exploring the characteristics of data and the most valuable are their own literal values.

We have performed some preliminaryexperiments on benchmark datasets in OAEI 2010 and 2011. The first dataset is a small real dataset, which includes two collections of RDF data files concerning persons (denoted by Person1 and Person2, respectively) and one collection about restaurants. OAEI 2010 organizers provided reference mappings for each collection, where each mapping contains two URIs from different data files that denote the same person or restaurant. The goal of our evaluation on the dataset is twofold. First, we want to test various values for the parameters in our approach and apply the best ones to the experiments. Second, we can compare the results obtained with other systems on the same dataset.

We compared the results of our approach (Relational) with other four entity/coreference resolution systems, namely ASMOV, CODI, LN2R and RiMOM, which also submitted their results on the same dataset to OAEI. ASMOV and CODI employed similarity-based matchers to obtain coreferent URIs and performed logical inference to remove inconsistent results. LN2R integrated a knowledge-based matcher to find semantically coreferent URIs and adopted a similarity propagation algorithm to generate similarities. RiMOM is a purely similarity-based system, which integrated many matchers to exploit a range of characteristics for both concepts and instances. All of them can only deal with pairwise instances, which are precisely called instance matching systems.

The comparison results on F-Measure is depicted in Fig. 6. From the bar graph, we can observe that our approach achieved the best F-Measure in average on the dataset. In particular, our Precision result is quite good , because we extracted a sensible number of evidences for the resolution process.

Fig. 6. F-Measure comparison among approaches on the benchmark test



In the second dataset, we compared our Relational approach with AgreementMaker, SERIMI, and Zhishi.Links, and that results were obtained from OAEI 2011 Instance Matching Campaign. AgreementMaker and Zhishi.Links are approaches that can be used only in one domain. Table 2 shows the comparative results with our approach. According to results, we got very much higher precision and recall if compared with other systems. Relational obtained good performance specifically on D4 and D5 datasets.

ISBN 978-85-61115-09-8

Table 2. Ourrelational approach compared with other systems results on OAEI2011 dataset

Dataset	Relational			Agree.Maker			Zishi.Links			SERIMI		
	PR	RC	F1	PR	RC	F1	PR	RC	F1	PR	RC	F1
D1	0.98	0.97	0.97	0.79	0.61	0.69	0.92	0.91	0.92	0.69	0.67	0.68
D2	0.97	0.95	0.96	0.84	0.67	0.74	0.90	0.93	0.91	0.89	0.87	0.88
D3	1.0	0.96	0.98	0.98	0.80	0.88	0.97	0.97	0.97	0.94	0.94	0.94
D4	0.98	0.95	0.96	0.88	0.81	0.85	0.90	0.86	0.88	0.92	0.90	0.91
D5	0.97	0.96	0.96	0.87	0.74	0.80	0.89	0.85	0.87	0.92	0.89	0.91
D6	1.0	1.0	1.0	0.97	0.95	0.96	0.93	0.92	0.93	0.93	0.91	0.92
D7	1.0	1.0	1.0	0.90	0.80	0.85	0.94	0.88	0.91	0.79	0.81	0.80
H-mean	0.97	0.97	0.97	0.92	0.80	0.85	0.93	0.92	0.92	0.89	0.88	0.89

We have performed someother preliminary tests with large datasets, more specifically with some datasets from BTC 2012 (Billion Triples Challenge) (HARTH, A., 2012), respectively RESTL and Freebase. These datasets contains approximately 122 million RDF n-quads triples. Firstly, we remove provenance information, duplicate triples, RDF blank nodes as well as reification statements. In the first round, it was generated a total of 675,244 *sameAs* links. The precision achieved was of 82% and the recall was of 75%. In the second round approximating 1 million links was generated, but the precision has dropped to 74% and the recall was of 68%. After all, the memory capacity was one of the big barriers we had to face, which showed a major drawback of the approach. Although the sparse nature of the matrix and tensor, we still had the problem of scale, dealing with millions of entities at the same time. Even so, we believe that the relational learning approach could result in the selection of the most likely mappings. Although, it is important to note that conducting more experiments is needed to deepen the discussion.

Our work has achieved good results with small real datasets and we are optmistic to perform more experiments with very large datasets employing the use of parallelization techniques as Map-Reduce. We believe that, as with any method based in machine learning addressing, as more data we have better can be the results.

Even with the different solutions there are still many open questions that deserve attention of researchers, e.g., take into account problems with unstructured, missing or erroneous data; use inter-relationships between entities; employ large-scale treatment; take into account for changes over time; perform resolution at query time. . Surveys about the topic can be found in (FERRARA, NIKOLOV and SCHARFFE, 2011; KÖPCKE and RAHM, 2010)

Referências

BIZER, C., HEATH, T., BERNERS-LEE, T. “Linked Data - The Story So Far.” *International Journal on Semantic Web and Information Systems* 5 (3), 2009 1–22 p.

BÖHM, C., DE MELO, G., NAUMANN, F., WEIKUM, G. “LINDA: Distributed Web-of-Data-Scale Entity Matching.” In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2104–8. CIKM ’12. New York, NY, USA: ACM, 2012.

COSTA, G. A., PARENTE DE OLIVEIRA, J. M. “A Relational Learning Approach for Collective Entity Resolution in the Web of Data.” In *Proceedings of the Fifth International Workshop on Consuming Linked Data*. Riva del Garda (IT): CEUR-WS, 2014. (a). To appear.

ISBN 978-85-61115-09-8

COSTA, G. A., PARENTE DE OLIVEIRA, J. M. “Large-Scale Entity Resolution for Semantic Web data Integration.” In *13th IADIS International Conference WWW/Internet*. Porto (PT): *Proceedings of the 13th International Conference WWW/Internet*, 2014. (b). To appear.

ELMAGARMID, A. K., IPEIROTIS, P.G., VERYKIOS, V.S. “Duplicate Record Detection: A Survey.” *IEEE Transactions on Knowledge and Data Engineering* 19 (1). 2007, 1 –16 p.

FELLEGI, I., SUNTER, A. “A Theory for Record Linkage.” *Journal of the American Statistical Association* 64 (328), 1969, 1183–1210 p.

FERRARA, A., NIKOLOV, A., SCHARFFE, F. “Data Linking for the Semantic Web.” *International Journal on Semantic Web and Information Systems* 7 (3), 2011, 46–76 p.

GETOOR, L. “Link Mining: A New Data Mining Challenge.” *SIGKDD Explorations*, 2003, 1–6 p.

GETOOR, L., DIEHL, C. P. “Link Mining: A Survey.” *SigKDD Explorations Special Issue on Link Mining,* 2005.

HEATH, T., BIZER, C. "*Linked Data: Evolving the Web Into a Global Data Space"*. Morgan & Claypool Publishers, 2011.

HOGAN, A., POLLERES, A., UMBRICH, J., ZIMMERMANN, A. “Some Entities Are More Equal than Others: Statistical Methods to Consolidate Linked Data.” In *Proceedings of the Workshop on New Forms of Reasoning for the Semantic Web: Scalable & Dynamic (NeFoRS2010)*. Heraklion, Greece, 2010.

HOGAN, A., ZIMMERMANN, A., UMBRICH, J., POLLERES, A., DECKER, S. “Scalable and Distributed Methods for Entity Matching, Consolidation and Disambiguation over Linked Data Corpora.” *Web Semantics: Science, Services and Agents on the World Wide Web* 10 (0), 2012, 76–110 p.

HU, W., QU, Y. Z., SUN, X. Z. “Bootstrapping Object Coreferencing on the Semantic Web.” *Journal of Computer Science and Technology* 26 (4), 2011, 663–675 p.

IOANNOU, E., PAPAPETROU, O., SKOUTAS, D., NEJDL, W. “Efficient Semantic-Aware Detection of near Duplicate Resources.” In *Proceedings of the 7th International Conference on The Semantic Web: Research and Applications - Volume Part II*, 136–50. ESWC’10. Berlin, Heidelberg: Springer-Verlag, 2010.

JAIN, P., HITZLER, P., SHETH, A. P., VERMA, K., YEH, P. Z. “Ontology Alignment for Linked Open Data.” In *Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part I*, 402–17. ISWC’10. Berlin, Heidelberg: Springer-Verlag, 2010.

JEAN-MARY, Y. R., SHIRONOSHITA, E. P., KABUKA, M. R. “Ontology Matching with Semantic Verification.” *Web Semant.* 7 (3), 2009, 235–251 p.

EUZENAT, J., SHVAIKO, P. "*Ontology Matching"*. Berlin Heidelberg (DE): Springer-Verlag, 2007.

KIM, H. S., LEE, D.. “HARRA: Fast Iterative Hashed Record Linkage for Large-Scale Data Collections.” In *Proceedings of the 13th International Conference on Extending Database Technology*. EDBT ’10. New York, NY, USA: ACM, 2010, 525–36 p.

KOLB, L., THOR, A., RAHM, E. “Load Balancing for MapReduce-Based Entity Resolution.” In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*. ICDE ’12. Washington, DC, USA: IEEE Computer Society, 2012a, 618–629 p.

KOLB, L., THOR, A., RAHM, E. “Multi-Pass Sorted Neighborhood Blocking with MapReduce.” *Computer Science - Research and Development* 27 (1), 2012b, 45–63 p.

ISBN 978-85-61115-09-8

KÖPCKE, HANNA, RAHM, E. “Frameworks for Entity Matching: A Comparison.” *Data Knowl. Eng.* 69 (2), 2010, 197–210 p.

NGOMO, A. C., AUER, S. “LIMES A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data.” In *IJCAI*, 2312–17. IJCAI/AAAI, 2011.

NGUYEN, K., ICHISE, R., LE, H. B. “Learning Approach for Domain-Independent Linked Data Instance Matching.” In Beijing, China, 2012.

NICKEL, M., TRESP, V., KRIEGEL, H. P. “A Three-Way Model for Collective Learning on Multi-Relational Data.” In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, edited by Lise Getoor and Tobias Scheffer. ICML ’11. New York, NY, USA: ACM, 2011, 809–816 p. (a)

NICKEL, M., TRESP, V., KRIEGEL, H. P., “Factorizing YAGO: scalable machine learning for linked data,” in *Proceedings of the 21st international conference on World Wide Web*, New York, NY, USA, 2012, pp. 271–280. (b)

NIU, X., RONG, S., ZHANG, Y., WANG, H. “Zhishi.links Results for OAEI 2011.” In *OM*, edited by Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz. Vol. 814. CEUR Workshop Proceedings. CEUR-WS.org, 2011.

PAPADAKIS, G., IOANNOU, E., PALPANAS, T., NIEDEREE, C., NEJDL, W. “A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces.” *IEEE Transactions on Knowledge and Data Engineering* 25 (12), 2013, 265–282 p.

PAPADAKIS, G., KOUTRIKA, G., PALPANAS, T., NEJDL, W. 2014. “Meta-Blocking: Taking Entity Resolution to the Next Level.” *IEEE Transactions on Knowledge and Data Engineering* 26 (8): 1946–1960.

PAPADAKIS, G., DEMARTINI. G., FANKHAUSER, P., KÄRGER, P. “The Missing Links: Discovering Hidden Same-as Links Among a Billion of Triples.” In *Proceedings of the 12th International Conference on Information Integration and Web-Based Applications & Services*. iiWAS ’10. New York, NY, USA: ACM, 2010, 453–460 p.

PAPADAKIS, G., IOANNOU, E., NIEDERÉE, C., PALPANAS, T., NEJDL, W. “Beyond 100 Million Entities: Large-Scale Blocking-Based Resolution for Heterogeneous Data.” In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. WSDM ’12. New York, NY, USA: ACM, 2012, 53–62 p.

RAIMOND, Y., SUTTON, C., SANDLER, M. “Automatic Interlinking of Music Datasets on the Semantic Web.” In , edited by Linked Data on the Web Workshop, 2008.

RETTINGER, A., LÖSCH, U., TRESP, V, D’ AMATO, C., FANIZZI, N. “Mining the Semantic Web - Statistical Learning for Next Generation Knowledge Bases.” *Data Mining and Knowledge Discovery* 24 (3), 2012, 613–662 p.

RONG, S., NIU, X., XIANG, E. W., WANG, H., YANG, Q., YU, Y.. “A Machine Learning Approach for Instance Matching Based on Similarity Metrics.” In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*. ISWC’12. Berlin, Heidelberg: Springer-Verlag, 2012, 460–475 p.

SHI, L., BERRUETA, D., FERNÁNDEZ, S., POLO, L., FERNAŃDEZ, S. “Smushing RDF Instances: Are Alice and Bob the Same Open Source Developer?” In *Personal Identification and Collaborations: Knowledge Mediation and Extraction*. Vol. 403. Karlsruhe, Germany: CEUR-WS, 2008.

SLEEMAN, J., AND FININ, T. “Learning Co-Reference Relations for FOAF Instances.” In *ISWC Posters&Demos*, Vol. 658. CEUR Workshop Proceedings. CEUR-WS.org, 2010.

ISBN 978-85-61115-09-8

SONG, D., HEFLIN, J. “Automatically Generating Data Linkages Using a Domain-Independent Candidate Selection Approach.” In *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*. ISWC’11. Berlin, Heidelberg: Springer-Verlag, 2011, 649–664 p.

TRESP, V., BUNDSCHUS, M., RETTINGER, A., HUANG, Y. “Towards Machine Learning on the Semantic Web.” In *Uncertainty Reasoning for the Semantic Web*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, 282–314 p.

VOLZ, J., BIZER, C, GAEDKE, M., KOBILAROV, G. “Discovering and Maintaining Links on the Web of Data.” In *Proceedings of the 8th International Semantic Web Conference*. ISWC ’09. Berlin, Heidelberg: Springer-Verlag, 2009, 650–665 p.

WINKLER, W. E. *Overview of Record Linkage and Current Research Directions*. 2006-2. Research Report Series. Washington, DC: Statistical Research Division, U.S. Census Bureau, 2006.

YILMAZ, Y. K., CEMGIL, A.-T, and SIMSEKLI, U. “Generalised Coupled Tensor Factorisation,” in Proceedings of Neural Information Processing Systems (NIPS), Granada, SPAIN, 2011, pp. 2151–2159.

KOLDA T. G., BADER B. W. "Tensor Decompositions and Applications". SIAM Rev. 2009; pp. 455–500.

HARTH, A. "Billion Triples Challenge Data Set". 2012. Avaiable at http://km.aifb.kit.edu/projects/btc-2012/

ISBN 978-85-61115-09-8

Mapeando Dados Governamentais com uma Ontologia de Organizações

Lucas B. R. da Fonseca[1]
lfonseca@inf.ufes.br

Carlos L. B. Azevedo[1,2]
clbazevedo@inf.ufes.br

João Paulo A. Almeida[1]
jpalmeida@ieee.org

[1]*Núcleo de Estudos em Modelagem Conceitual e Ontologias (NEMO), UFES, Vitória, ES*
[2]*Services, CyberSecurity and Safety Research Group (SCS), University of Twente, Holanda*

Resumo
Em 18 de novembro de 2011 foi sancionada a Lei de Acesso à Informação (Lei nº 12.527/2011) que regula o acesso às informaçõesmantidas pelo governo nas esferas federal, estadual e municipal. Com essa lei em vigor, um maior volume de dados de caráter público passou a ser disponibilizado na Internet. Apesar da maior disponibilidade, os dados fornecidospelas várias organizações governamentaispossuem formatos diferentes e advêm de fontes heterogêneas e não integradas, o que dificulta a utilização destes pelos cidadãos e sua apropriação para reuso em sistemas computacionais. Uma estratégia recente recomendada por órgãos de padronização como o W3Cprevê o uso de ontologias e tecnologias semânticaspara integrar e disponibilizar esses dados. Este artigo relata um estudo de caso na integração de dados governamentais no domínio depublicação de informações sobre organizações e estruturas organizacionais utilizando a W3C ORG Ontology. No estudo, foramrecuperados dados publicados em seu formato corrente, realizado um mapeamento destes dados para aontologia de referência da W3C e publicados os dados integrados em RDF em conformidade com a ontologia. O artigo apresenta a abordagem adotada,seus benefícios e discute as dificuldades encontradas assim comoas lições aprendidas em cada uma das fases do processo.

Palavras-chave: Web Semântica, Dados Ligados, Ontologias, Integração de dados.

Abstract
On November 18 2011, Brazil passed its Freedom of Information Law (n. 12.527/2011) regulating access to government information at all levels (federal, state and municipal). With the enactment of this law, public government data has been increasingly made available on the Internet. Despite its availability, data published by government organizations often adopts ad hoc formats and originates from heterogeneous sources with little or no integration, creating barriers for data consumption.In order to address this challenge, several standards bodies such as W3C have proposed the use of ontologies and semantic technologies.This paper reports on case study on data integration in the domain of organizational structures using the W3C ORG Ontology. We discuss the approach that has been employed along with its potential benefits. Lessons learned throughout the integration process are discussed.

Key Words: Semantic Web, Linked Data, Ontologies, Data integration.

Introdução

ISBN 978-85-61115-09-8

Em 18 de novembro de 2011 foi sancionada a Lei de Acesso à Informação (Lei nº 12.527/2011) que regula o acesso às informaçõesmantidas pelo governo nas esferas federal, estadual e municipal. Essa lei constituiu um avanço para a democratização da informação pública, com um grande volume de dados governamentais passando a ser disponibilizado na Internet. O objetivo é proporcionar ao cidadão maiorvisibilidade às açõesde governo, melhor acesso aos serviços públicos emaior controle das contas públicas através de mecanismos de transparência.

Exemplos de iniciativas federais em resposta às demandas de acesso à informação pública incluem oPortal de Dados Abertos (disponível em http://dados.gov.br) e o Portal da Transparência (disponível em http://www.portaltransparencia.gov.br). Atualmente, os dados que são publicados por estes portais possuem formatos diferentes e advêm de fontes heterogêneas e não integradas, o que dificulta a utilização destes pelos cidadãos e sua apropriação para reuso em sistemas computacionais. A publicação dos dados é usualmente feita através de arquivos não estruturados (tais como planilhas e arquivos PDF), dificultando a leitura,integração e análise automatizadas.

Para lidar com as limitações das abordagens *ad hoc* para publicação de dados, o W3C (*World Wide Web Consortium*)vem recomendando um conjunto de boas práticas para publicar dados de forma estruturada e interligada (BIZER; HEATH; BERNERS-LEE, 2009). Esta abordagem preconiza o uso de vocabulários compartilhados que são representados através de *ontologias* formalizadas em linguagens de representação de conhecimento concebidas para a Web e objetivam o melhor compartilhamento, consumo e integração de dados.

Este artigorelata um estudo de caso empregando a abordagem de integração baseada em ontologias utilizando bases de dados governamentais. Foram mapeados os diversos dados em seus formatos correntes para uma ontologia de organizações padronizada pelo W3C (denominada *ORG Ontology*(REYNOLDS, 2014)). A *ORG Ontology*foi criada com o intuito de prover um modelo genérico para publicação de informações sobre organizações e estruturas organizacionais, incluindo organizações governamentais. A *ORG Ontology*permite descrever a estrutura de organizações, bem como as pessoas envolvidas, informações sobre a localização das organizações e seus históricos (como fusões e mudanças de nomes) (REYNOLDS, 2014).

Em especial, busca-se nesse trabalho: (i) avaliar as dificuldades existentes para integrar os dados abertos disponibilizados pelo governo brasileiro, incluindo prováveis dificuldades de interpretação desses dados, com o uso de uma ontologia recomendada pelo W3C; (ii) avaliar a abrangência e adequação da ontologia aos dados governamentais brasileiros e; (iii) considerar os benefícios e limitações do uso da abordagem em um caso real.

Este artigo está organizado da seguinte forma. A seção 1 aborda o referencial teórico relevante ao contexto deste trabalho. A seção 2 apresenta a abordagem para realização do mapeamento, desde da etapa conceitual atéa sua implementação. A seção 3 apresenta uma aplicação que usa os resultados do mapeamento visando demonstrar o consumo dos dados em conformidade com a *ORG Ontology*. A seção 4 apresenta as discussões acerca do mapeamento, mostrando as dificuldades encontradas e os benefícios das tecnologia semânticas empregadas. Por fim, a última seção apresenta as conclusões do trabalho, limitações e propostas de trabalhos futuros.

1 Referencial Teórico

O mapeamento apresentado neste trabalho foi realizado com base em padrões do W3C, e envolve diretamente linguagens como RDF (*Resource Description Framework*) e SPARQL (*SPARQL Protocol and RDF Query Language*) dentro de contextos como a Web

ISBN 978-85-61115-09-8

Semântica, Dados Ligados (*Linked Data*) e Ontologias. Essa seção apresenta estes elementos para contextualizar o trabalho.

1.1 Dados Ligados e Ontologia

Berners-Lee (2006) define um conjunto de regras para a publicação de dados na Web, chamados de Dados Ligados (*Linked Data*). A intenção é que dados publicados seguindo esse conjunto de regras sejam unificados em único espaço global de dados. As regras propostas são: (1) Usar URI como nome para as coisas; (2) Usar URIs HTTP para que as pessoas possam procurar por estes nomes; (3) Fornecer informações úteis na recuperação de URIs, usando os padrões RDF e SPARQL; (4) Incluir links para outros URIs, para que seja possível descobrir mais informações.

Estas regras fornecem os princípios básicos para a publicação e conexão de dados através da Web. Para propiciar distinção semântica entre os dados publicados, possibilitando integração desses dados com dados de outras fontes, utilizam-se ontologias.

As ontologias mais comuns na Web possuem uma taxonomia e um conjunto de regras de inferência(BERNERS-LEE; HENDLER; LASSILA, 2001). Taxonomias definem classes de objetos e as relaçõesentre elas. As regras de inferência permitem melhorar a qualidade da análise dos dados, através da descoberta de novas relações automaticamentesobre o conteúdo existente, além de detectar possíveis inconsistência neles.

Para representar uma ontologia na Web, utiliza-se a Linguagem de Ontologia para Web (OWL - *WebOntology Language*) que é uma extensão de RDF *Schema.*A linguagem OWL provê maior capacidade de descrição de classes e propriedades em relação ao RDF, incluindo a expressão de relações entre classes, restrições de cardinalidade, características de propriedades e classes enumeradas (MCGUINNESS; HARMELEN, 2004).

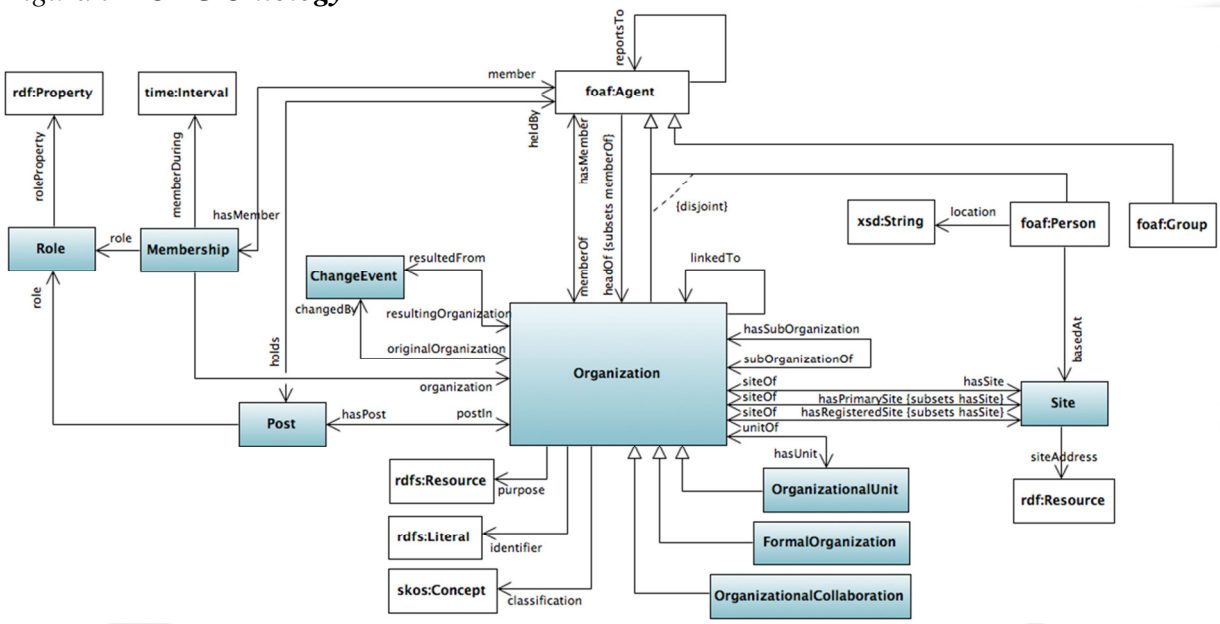
1.2 ORG Ontology

A *ORG Ontology* (*The Organizational Ontology*) é uma recomendação W3C, descrita em (REYNOLDS, 2014), projetada para permitir a publicação de informações sobre as organizações e estruturas organizacionais, incluindo organizações não-governamentais. A ontologia descreve a estrutura da organização, bem como as pessoas envolvidas nessa estrutura, informações sobre a localização da organização e o histórico da organização (como fusões e mudanças de nomes). A *ORG Ontology* possui um conjunto de conceitos bem estruturados e definidos, possibilitando que outros indivíduos entendam e os utilizem. Elatambém reutiliza conceitos de outras ontologias usadas em larga escala (p. ex., FOAF, SKOS e *vCard*).

A *ORG Ontology* fornece um modelo genérico e reutilizável. Esse modelo pode ser estendido ou especializado para uso em situações particulares. A Figura 1 apresenta um diagrama (não normativo) da*ORG Ontology*.

ISBN 978-85-61115-09-8

Figura 7 **– *ORG Ontology***



Fonte: (REYNOLDS, 2014)

Conforme a Figura 1, **Organização**(*Organization*) representa um conjunto de pessoas organizadas em uma comunidade ou outra estrutura social, comercial ou política. Uma **Organização** pode ser especializada em uma **Unidade Organizacional** (*OrganizationalUnit*), em uma **Organização Formal** (*FormalOrganization*) ou em uma **Colaboração Organizacional**(*OrganizationalCollaboration*). **Unidades Organizacionais**são entidades (como departamentos ou unidades de suporte) que pertencem a uma **Organização**e que não podem ser consideradas como uma entidade legal de direitos próprios. **Organizações Formais** são organizações reconhecidas mundialmente (como corporações, instituições de caridade, governo ou igreja) através de jurisdições legais, direitos e responsabilidades associadas. **Colaboração Organizacional**define um tipo de colaboração entre duas ou mais **Organizações** como um projeto. Ela atende aos critérios para ser uma organização na medida em que tem uma identidade e definição de propósito independente de seus membros em particular, mas não é nem uma entidade jurídica formalmente reconhecida nem uma unidade organizacional dentro de uma organização maior. Exemplos de **Colaboração Organizacional**são projetos conjuntos entre várias organizações ou consórcios de organizações, como os formados para licitações de grandes obras.**Adesão** (*Membership*) de uma **Organização** representa o **Papel** (*Role*) que o **Agente** (*Agent*) tem na **Organização** durante um **Intervalo** (*Interval*) de tempo. **Agente** detém de um **Cargo** (*Post*) e pode ser uma **Organização** (exceto a qual ele é membro), uma **Pessoa** (*Person*) ou um **Grupo** (*Group*). **Pessoa** e **Organização** têm **Endereço** (*Site*). **Organização** pode ser alterada por **Eventos de Mudança** (*ChangeEvent*) que representam eventos que resultaram em grandes alterações para uma **Organização** como uma fusão ou reestruturação completa. Além disso, a **Organização** possui um **identificador** (*identifier*), um **propósito** (*purpose*), uma **classificação** (*classification*), essa última podendo variar dentro de algum esquema de classificação, e um ou vários **Endereços**(REYNOLDS, 2014).

2Mapeamento

Esta seção aborda as etapas para o mapeamento dos dados contidos nas bases governamentais selecionadas para dados estruturados, de acordo com a *ORG Ontology*. Para tal foi utilizada uma linguagem de mapeamento de dados relacionais para RDF, a linguagem

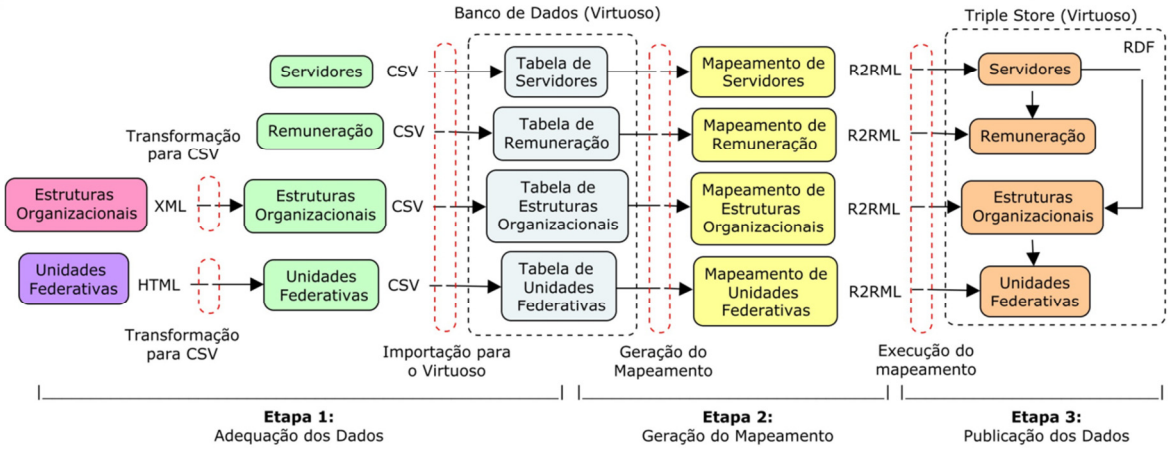
ISBN 978-85-61115-09-8

R2RML (DAS; SUNDARA; CYGANIAK, 2012). A escolha das bases de dados baseou-se na relevância dos dados e em sua representatividade em relação a *ORG Ontology,* verificando se esses dados possuem interesse público, se podem ser mapeados aos termos do vocabulário da *ORG Ontology* e se, em conjunto, representam uma parte significativa de conceitos da ontologia. Foram escolhidos dados advindos de diversas fontes governamentais para verificar a integração existente entre os dados.Todas as bases de dados selecionadas pertencem a órgãos do governo federal. Com isso, busca-se verificar tanto a integração dos dados disponibilizados pelo governo quanto a abrangência da ontologia no escopo analisado.

Foram utilizadas as seguintes bases de dados: (i) **Servidores Civis (Servidores e Remuneração),** que apresenta dados em formato *Comma-Separated Values* (CSV) sobre cargo, função, situação funcional e remuneração dos servidores civis; (ii)**Estruturas Organizacionais,** que apresenta dados em formato XML contendo informações organizacionais do Poder Executivo Federal, (Administração Direta, Autarquias e Fundações), tais como: nomes, códigos e endereços de órgãos públicos e suas subdivisões administrativas e; (iii)**Catálogo de Unidades Federativas,**que apresenta dados em formato HTML contendo as 27 Unidades Federativas em acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE).

Para a realização do mapeamento, foram necessárias várias etapas para ter como resultado final um Grafo RDF que segue os princípios dos Dados Ligados e semântica de dados bem definidas. Essas etapas são ilustradas na Figura 2 e são descritas nas subseções a seguir.

*Figura 8***– *Etapas do Mapeamento***



Fonte: Elaborado pelo autor

2.1 Etapa de Adequação dos Dados

Inicialmente, com a verificação de que as bases escolhidas retornam formatos diferentes, foi necessária a transformação de todas as bases de dados para um formato único. O formato CSV foi escolhidodevido a maior quantidade dos dados publicados no escopo do trabalho já estarem neste formato, para que não fosse necessária alteração de formato dos dados sobre Servidores e Remuneração. Os dados de Estruturas Organizacionais foram transformados de XML para CSV, com o auxílio da ferramenta *Google Refine*[1], que transforma dados XML em uma tabela, permitindo salvar no formato CSV. Como a quantidade de dados sobre Unidades Federativas era pequena, a transformação foi feita manualmente copiando os dados e inserindo-os em um arquivo CSV.

[1]https://code.google.com/p/google-refine/

ISBN 978-85-61115-09-8

Em seguida, todos os dados foram importados para obanco de dados de triplas, convertendo os dados em CSV para tabelas lógicasatravés de uma opção presente no banco que permite realizar esse tipo de operação. O banco de dados de triplasescolhido foi o *Virtuoso*[2]. A escolha se baseou nos fatos de o banco permitir armazenar tanto os dados originais quanto os dados em RDF resultantes e, especialmente, possuir suporte a linguagem de mapeamento R2RML, utilizada no mapeamento dos dados com a ontologia.

2.2 Etapa de Geração do Mapeamento

Na etapa de geração do mapeamento, um mapeamento conceitual foi realizado de forma a classificar os elementos das bases de dados com os conceitos da *ORG Ontology*. A análise foi feita com base nas definições dos conceitos da ontologia e dos dados das bases de dados, ou seja, relacionando os elementos das bases com o conceito cuja definição fosse a mais apropriada. A Tabela 1 apresenta um fragmento do mapeamento conceitual, da base de dados de Servidores Civis com os conceitos da *ORG Ontology*.

Tabela 3 –Fragmento do mapeamento conceitual sobre as colunas da base de dados de Servidores Civis com os termos da ORG Ontology

Servidores	Conceito da Org Ontology
Servidor	Pessoa (*Person*)
Órgão de Lotação	Organização Formal (*FormalOrganization*)
Órgão de Exercício	Organização Formal (*FormalOrganization*)
Cargo	Cargo (*Post*)
Atividade	Cargo (*Post*)

Fonte: Elaborado pelo autor

Seguindo a Tabela 1, os elementos presentes na linha Servidor representam pessoas que mantém vínculos de trabalho com entidades governamentais. Esses elementos foram mapeados ao conceito de **Pessoa** (*Person*) da *ORG Ontogy*. Já os elementos presentes na segunda e terceira linha, Órgão de Lotação e Órgão de Exercício, apesar de possuírem significados diferentes em relação ao Servidor (o primeira representa onde o servidor está lotado e o segundo onde ele exerce suas atribuições), representam a entidade Órgão Governamental, e, por isso, são mapeados ao conceito **Organização Formal** (*Formal Organization*) da *ORG Ontology*. Caso semelhante a este, com mapeamento de dois elementos da base de dados a um único conceito ocorre no mapeamento sobre Cargo e Atividade (últimas duas linhas da Tabela 1), onde o primeiro representa um conjunto de atribuições inerentes ao agente público e o segundo um conjunto de atribuições inerentes ao exercício de funções especiais, como chefia e assessoramento. No entanto, esta distinção não existe na *ORG Ontology*. Dessa forma, o mapeamento foi feito ao conceito da ontologia mais adequado a ambos, no caso, ao conceito **Cargo** (*Post*).

Após a realização do mapeamento conceitual, foram criadas representações mais formais dos mapeamentos conceituais, de forma a se adequar a linguagem R2RML. Para representar o padrão de mapeamento aqui descrito, foi criada uma representação visual, em que um diagrama representa um *template* de um grafo que será gerado na saída do mapeamento. Esse *template* é parametrizado com variáveis que representam dados originários das bases de dados de entrada.São representados nesta notação os seguintes elementos:**Classes das Ontologias (laranja**), que representam os termos dos vocabulários das ontologias contento o prefixo da ontologia mais o termo do vocabulário, denotados na

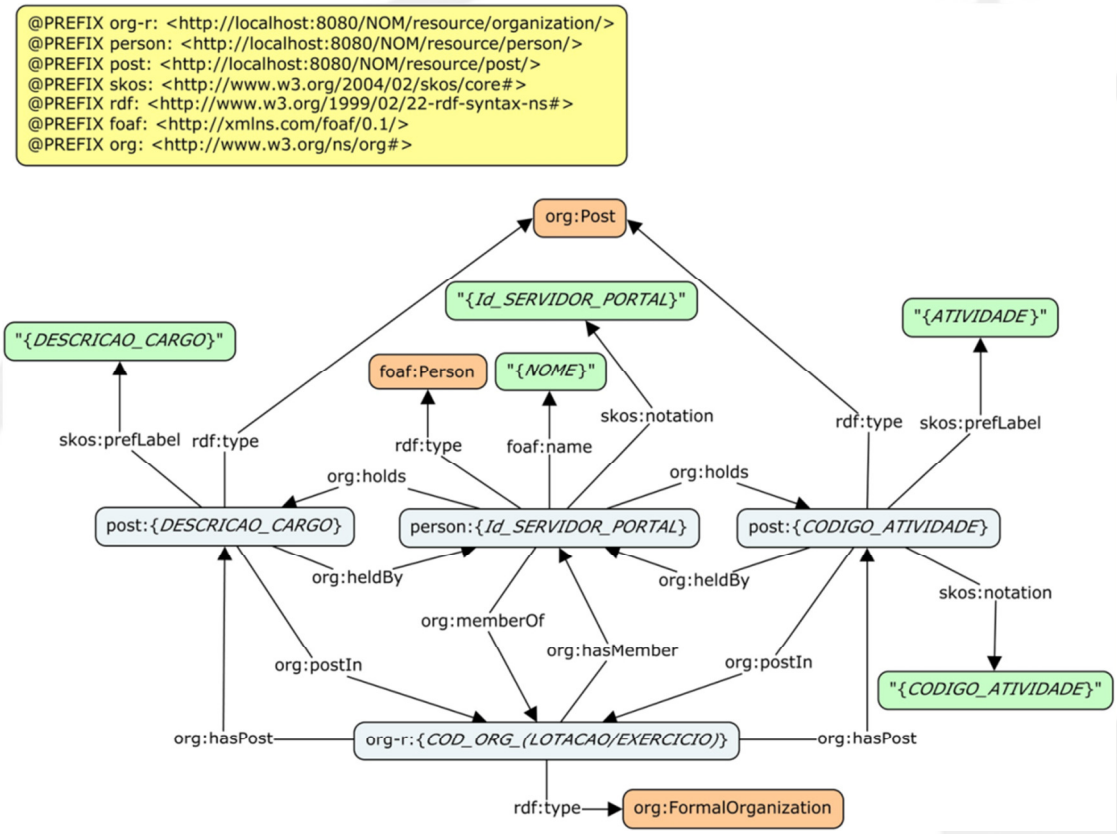
[2]Foi utilizado o banco de dados Virtuoso v 6.1.7. Disponível em: http://sourceforge.net/projects/virtuoso/files/virtuoso/6.1.7/

ISBN 978-85-61115-09-8

forma <Prefixo>:<Termo do vocabulário>; **Entidades (azul)**, que representam um URI contendo o prefixo criado para a base de dados mais uma variável dada pelo valor da coluna mapeada, denotados na forma <Prefixo>:<{*Variável*}> e; **Literais (verde)**, que representam o valor da coluna mapeada, denotados na forma <"*Variável*">.

A Figura 3 apresenta o fragmento do mapeamento da relação entre servidores, cargos, órgãos de lotação e exercício.

Figura 9–Fragmento referente ao mapeamento da relação entre servidores, cargos e organizações



Fonte: Elaborado pelo autor

Conforme a Figura 3, cada servidor é representado pelo seu identificador na base de dados (Id_SERVIDOR_PORTAL). Servidores são um tipo (*rdf:type*) de Pessoa (*foaf:Person*) e possuem um nome (*foaf:name*) representado pela coluna "NOME" e o identificador único (*skos:notation*) representado pela coluna "Id_SERVIDOR_PORTAL". Servidores têm (*org:holds*) cargo ou atividade e são membros (*org:memberOf*) de um Órgão de Lotação e Exercício. Como a base de dados não possui identificadores para cargos, seu URI e seu rótulo preferencial (*skos:prefLabel*) são criados utilizando a coluna "DESCRICAO_CARGO". Já para atividades, temos que o URI utiliza o código (CODIGO_ATIVIDADE) e seu rótulo preferencial (*skos:prefLabel*) é representado pela coluna "ATIVIDADE". Além disso, Atividade possui o identificador único (*skos:notation*) representado pela coluna "COGIDO_ATIVIDADE". Ambos cargos e atividades são mapeados como um mesmo tipo (*rdf:type*) Cargo (*org:Post*) e são cargos (*org:postIn*) dentro de um órgão de lotação e exercício. Tanto Órgão de Lotação quanto de Exercício possuem membros (*org:hasMember*) e cargos e atividades (*org:hasPost*).

ISBN 978-85-61115-09-8

2.3 Etapa de Publicação dos Dados

Na etapa de publicação dos dados, foi realizada a codificação dos dados na linguagem de mapeamento R2RML com base no mapeamento conceitual e nos princípios dos Dados Ligados. A publicação, no entanto, é feita apenas localmente, com o intuito de mostrar, posteriormente, os resultados obtidos com a execução do mapeamento. A Listagem 1 apresenta o código referente a parte do mapeamento da base de dados de Servidores Civisna linguagem R2RML.

Listagem 1 – Código em R2RML referente a parte do mapeamento da base de dados de Servidores Civis

```
DB.DBA.TTLP ('
@prefix rr: <http://www.w3.org/ns/r2rml#> .@prefix org: <http://www.w3.org/ns/org#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
<http://localhost:8080/NOM/resource#TriplesMapPerson>
  a rr:TriplesMap;
  rr:logicalTable
  [ rr:tableSchema "CSV"; rr:tableOwner "DBA"; rr:tableName "Agentes_csv" ];
  rr:subjectMap
  [ rr:template "http://localhost:8080/NOM/resource/person/{Id_SERVIDOR_PORTAL}";
    rr:class foaf:Person;
    rr:graph <http://localhost:8080/NOM/graph#>;];
rr:predicateObjectMap
  [rr:predicate foaf:name;
rr:objectMap [ rr:column "NOME" ];];
   rr:predicateObjectMap [
     rr:predicate org:memberOf;
     rr:objectMap [
     rr:parentTriplesMap <http://localhost:8080/NOM/resource#TriplesMapOrganization>;
        rr:joinCondition [rr:child "COD_ORG_LOTACAO";
            rr:parent "COD_ORG_LOTACAO";];];].
', 'http://temp/person', 'http://temp/person');
```

Fonte: Elaborado pelo autor

No código, inicialmente são definidos os prefixos dos vocabulários RDF. O prefixo *rr* representa o vocabulário da linguagem R2RML. Em seguida, é definido um mapa de tripla como um tipo de mapa de triplas (*rr:TripleMap*). Após isso, é indicada a base de dados (*rr:logicalTable*) que será mapeada. Então, é definido o sujeito da tripla (*rr:subjectMap*) com um URI padrão (*rr:template*) usando um ou vários dados da base de dados, no caso, o identificador do servidor (Id_SERVIDOR_PORTAL). É definido então o tipo (*rr:class* que no resultado do mapeando vira um *rdf:type*) do URI e o grafo RDF (*rr:graph*) no qual será inserido as triplas RDF. Adiante, temos os mapeamentos de predicados e objetos (*rr:predicateObjectMap*) que serão ligados ao sujeito da tripla RDF. O predicado (*rr:predicate*) será um URI de algum vocabulário. E o objeto pode ser um literal que é uma coluna (*rr:column*) da base de dados ou um outro URI de outro mapa de triplas. Neste ultimo caso, é feita uma junção (*rr:joinCondition*) com a outra base de dados através da coluna da base atual (*rr:child*) com a coluna da base do outro mapa de triplas (*rr:parent*). Ele consiste em ligar o sujeito do mapa atual com o sujeito do outro mapa (*rr:parentTripleMap*) quando os valores das duas colunas forem iguais.Por fim, o mapeamento é criado em um grafo temporário (http://temp/person).

Após a criação de todos os grafos temporários (aqui representado somente o de pessoa), ambos são inseridos em um grafo geral (http://temp/mix) através de uma inserção via SPARQL, conforme o código abaixo. O resultado, então, é inserido no grafo (representado no exemplo pelo grafo http://localhost:8080/NOM/graph#) indicado no mapeamento do sujeito.Por fim, é realizada uma inserção via SPARQL, conforme a Listagem 2, para transformar os dados do grafo, indicado no mapeamento do sujeito, para um grafo RDF representado por um URI pré-definido (no caso, http://localhost:8080/NOM/resource#).

ISBN 978-85-61115-09-8

Listagem 2 – Transformação dos dados para um grafo RDF com inserção via SPARQL

```
sparql insertin graph <http://temp/mix> { ?s ?p ?o }
from<http://temp/person>where { ?s ?p ?o };
exec ('sparql' || DB.DBA.R2RML_MAKE_QM_FROM_G ('http://temp/mix'));
sparql insertin graph <http://localhost:8080/NOM/resource#> { ?s ?p ?o }
from<http://localhost:8080/NOM/graph#>where { ?s ?p ?o };
```

Fonte: Elaborado pelo autor

A base de dados de Servidores Civis possuía cerca de 735 mil linhas e foram utilizadas 15 colunas. A base de dados de remuneração possuía cerca de 570 mil linhas e foram utilizadas 6 colunas. A base de Estruturas Organizacionais possuía um pouco mais de 70 mil linhas e foram utilizadas 19 colunas. O mapeamento gerou um total de cerca de 9,5 milhões de triplas.

3Aplicação do Mapeamento e Uso dos Dados Mapeados

Seguindo os princípios dos Dados Ligados, uma aplicação local foi desenvolvida para demonstrar as potencialidades do mapeamento realizado e simular o uso dos dados publicados na Web. Nela é possível acessar os recursos através dos URIs que utilizam o protocolo HTTP, em um processo chamado *dereference*, que apresenta ao usuário algum conteúdo referente ao recurso acessado.

Na aplicação, o usuário consegue dereferenciar um URI qualquer e visualizar suas informações. A partir de um URI, o usuário pode navegar para os URIs relacionados ao URI original e acessar as suas informações, visualizando mais informações. A aplicação também permite que os termos das ontologias possam ser acessados e apresenta as suas definições.

AFigura 4mostra um exemplo de um dereferenciamento do URI `http://localhost:8080/NOM/resource/person/1001212`, pertencente a servidora “Ana Cristina Ribeiro Alvim”[3], que trás os elementos conectados diretamente ao URI.

*Figura 10***– *Exemplo de URI Dereferenciado***

Property	Value
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://xmlns.com/foaf/0.1/Person
http://www.w3.org/2004/02/skos/core#notation	1001212
http://xmlns.com/foaf/0.1/name	ANA CRISTINA RIBEIRO ALVIM
http://www.w3.org/ns/org#memberOf	http://localhost:8080/NOM/resource/organization/26411
http://www.w3.org/ns/org#memberOf	http://localhost:8080/NOM/resource/organizational-unit/CAMPUS%20BCENA%20-%20DIRETORIA%20ADMIN%20E%20PLANEJA
http://www.w3.org/ns/org#holds	http://localhost:8080/NOM/resource/post/CONTADOR
http://www.w3.org/ns/org#remuneration	http://localhost:8080/NOM/resource/remuneration/brl/2013/11/person/1001212
http://www.w3.org/ns/org#remuneration	http://localhost:8080/NOM/resource/remuneration/usd/2013/11/person/1001212

Fonte: Elaborado pelo autor

As representações descritas acima permitem que usuários da aplicação visualizem a informação requerida de forma mais fácil em relação ao formato original, além de permitir uma melhor compreensão das informações, dado os termos estão mapeados a uma ontologia de referência.

Consultas específicas podem ser realizadas através da linguagem SPARQL, sendo possível obter o resultado de consultas em diversos formatos (RDF, HTML, CSV, etc.). O

[3]A servidora específica foi escolhida aleatoriamente na base de dados. As informações dispostas neste trabalho foram recuperadas de dados publicados sob dados abertos governamentais, de acordo com a lei 12.527/2011 da República Federativa do Brasil.

ISBN 978-85-61115-09-8

formato de acesso também permite que os elementos sejam usados por aplicações externas. OBanco de Dados de Triplas disponibiliza um SPARQL *Endpoint* que pode ser acessado pelo usuário. Ele possibilita a realização de consultas de forma programática (através de código) e de forma visual interativa (com consultas através de grafos visuais). O SPARQL *Endpoint* possibilita ao usuário realizar consultas com inferência,ou seja, consultas que realizam raciocínio automático através de regras definidas para descobrir novas relações, i.e., relações não anteriormente mostradas entre os elementos do grafo.

Para exemplificar o uso de SPARQL com inferência, a Listagem 3 apresenta uma consulta que visa encontrar o total de membros da "Fundação Nacional de Saúde". Na primeira linha, temos definido o grafo que possui as regras de inferência usadas na consulta, no caso, o grafo da *ORG Ontology* (representado pelo URI http://www.w3.org/ns/org#). A *ORG Ontology* define em seu modelo diversas regras (como relações inversas e transitividade) que permitem tirar conclusões automáticas sobre os dados.

Listagem 3 – Exemplo de consulta SPARQL com inferência

```
DEFINE input:inference <http://www.w3.org/ns/org#>
PREFIX org: <http://www.w3.org/ns/org#>
SELECT COUNT(?person) WHERE {
?org org:hasMember ?person .
   ?org skos:prefLabel "FUNDACAO NACIONAL DE SAUDE" . }
```

Fonte: Elaborado pelo autor

Ao realizar a consulta, temos como resposta que a "Fundação Nacional de Saúde" possui um total 13254 servidores. Sem o uso de inferências, não seria possível encontrar este valor, dado que existe apenas elementos relacionados pelo predicado *memberOf* que são encontrados automaticamente através da relação inversa com o predicado *hasMember*definido na *ORG Ontology*.

4Discussões

A realização do mapeamento possibilitou melhorias ao uso e compreensão dos dados. Entretanto, algumas dificuldades para se chegar a essa situação foram encontradas, devido a problemas existentes nos dados recebidos e em sua interpretação. As subseções a seguir apresentam os problemas encontrados para a correta realização do mapeamento e o uso dos dados, assim como os benefícios advindos do uso dos dados mapeados e integrados.

4.1 Problema da Falta de Identificação Única

Um problema encontrado foi a falta de identificação única para as entidades. Esse problema faz com que seja difícil a correta identificação, relacionamento e separação dos dados. Isso faz com que dados iguais, porém apresentados diferentemente em sua forma sintática são classificados como dados diferentes. Da mesma forma, dados diferentes, porém apresentados com a forma sintática igual, podem ser classificados como o mesmo elemento.

Nesse trabalho, como não foi encontrado um identificador único para os dados, para remediar esse problema, foram utilizados identificadores sintáticos para realizar a identificação única dos elementos. Além disto, o uso de maiúsculas e minúsculas, assim como caracteres especiais também dificultou a identificação única dos elementos. Para tal, uma transformação dos dados com a retirada de caracteres especiais e com a modificação das bases para caracteres em maiúsculos também foi efetuada. Como exemplo, a Tabela 2 apresentaalguns cargos da base de dados de Servidores Civis que aparentemente são os mesmos, mas, por possuírem distinções sintáticas e não possuírem identificação única, são mapeados como entidades diferentes.

ISBN 978-85-61115-09-8

Tabela 4 - Exemplo de Cargos semelhantes sem identificadores da base de dados de Servidores Civis

CARGO
AGENTE DE TELEC E ELETRICIDADE
AGENTE DE TELEC ELETRICIDADE
AGENTE DE TELECOMUNI E ELETRICIDADE
AGENTE DE TELECOMUNIC E ELETRICIDADE

Fonte: Elaborado pelo autor

Esse problema afeta, por exemplo, a precisão de relatórios e estatísticas que poderiam ser gerados sobre esses dados.A falta de identificação única afeta também, em alguns casos, o mapeamento de determinadas entidades distintas, mas que possuem a mesma descrição e/ou nome descritivo. Para exemplificar essa situação, a Figura 5 apresenta o caso onde o nome "Coordenação de Planejamento" é usado para diversas organizações. Provavelmente trata-se de diferentes unidades organizacionais, dado o fato de pertencerem a órgãos diferentes. Apesar disto, não é possível saber ao certo se trata-se de uma coincidência de descrições ou de fato de entidades diferentes.

O uso de identificadores únicos resolveria os problemas mencionados nessa seção. Eles unificariam entidades semanticamente iguais independentemente de suas descrições e permitiriam a distinção de entidades com as mesmas descrições.

Figura 11 – Exemplo de problema da falta de identificadores no mapeamento de unidades organizacionais

Property	Value
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/ns/org#OrganizationalUnit
http://www.w3.org/2004/02/skos/core#prefLabel	COORDENACAO DE PLANEJAMENTO
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/36205
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/25000
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/40112
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/45203
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/20224
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/28000
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/26442

Fonte: Elaborado pelo autor

4.2 Problema de Falta deIntegração entre Diferentes Bases de Dados

Outro problema encontrado refere-se à falta de integração entre as diferentes bases de dados do governo. Esse problema faz com que seja difícil a correta identificação, separação e relacionamento dos dados entre as diversas bases. Entidades, quando possuem identificação em uma base específica, não possuem a identificação compartilhada com as demais bases do governo, de forma que não é possível precisar com certeza a identificação única e correta de elementos referenciados nas várias bases.

Nesse trabalho a identificação única se tornou especialmente complexa devido à falta de identificadores únicos verificada em dados recuperados de uma mesma base, conforme descrito na seção 4.1. A abordagem utilizada para remediar esses problemas foi a de

ISBN 978-85-61115-09-8

classificar os dados de acordo com as descrições textuais. Todos os dados foram transformados para terem caracteres em maiúsculos e foram retirados os caracteres especiais de sua forma sintática para diminuir a probabilidade de separações incorretas. Reitera-se porém que o trabalho pode ter classificado o mesmo elemento, porém apresentado diferentemente em sua forma sintática nas diversas bases como dados diferentes. Da mesma forma, dados diferentes, porém apresentados com as características sintáticas iguais podem ter sido incorretamente classificados como o mesmo dado. Reitera-se a complexidade dado que a representação em cada base pode apresentar distinções nos caracteres utilizados para um mesmo dado, como abreviações e siglas.

A Tabela 3, onde "Ministério da Fazenda" possui diversos códigos advindos das diversas bases de dados, exemplifica o problema da falta de identificação única apresentado.

Tabela 5 – Diferentes identificadores para Ministério da Fazenda nas bases de dados

NOME	CÓDIGO	CÓDIGO PAI
MINISTERIO DA FAZENDA	1929	26
MINISTERIO DA FAZENDA	118699	1929
MINISTERIO DA FAZENDA	118700	1929
MINISTERIO DA FAZENDA	118701	1929
MINISTERIO DA FAZENDA	118702	1929
MINISTERIO DA FAZENDA	118703	1929
MINISTERIO DA FAZENDA	118704	1929
MINISTERIO DA FAZENDA	118708	1929

Fonte: Elaborado pelo autor

Continuando no exemplo da Tabela 3, nesse caso, no mapeamento automático os órgãos descritos acima são mapeados como "Ministério da Fazenda". Entretanto, ao se realizar uma navegação manual nos dados abertos do governo (disponibilizador dos dados), é possível verificar, a exceção do primeiro elemento, que os outros órgãos descritos são órgãos vinculados ao Ministério da Fazenda e não o próprio Ministério da Fazenda.

Para a solução do problema acima descrito, a abordagem preferencial seria que os sistemas governamentais fossem integrados, de forma que elementos semanticamente iguais tivessem um mesmo identificador. Isso poderia ser realizado tanto com um código identificador quanto com uma identificação sintática única.

A integração dos dados possibilitaria um compartilhamento maior e interpretação melhor das informações, com o cruzamento de dados advindos de bases diferentes.

4.3 Problema da Falta de Expressividade na ORG Ontology

A *ORG Ontology* foi utilizada como ontologia de referência e para adicionar semântica aos termos mapeados nesse trabalho. A *ORG Ontology* abrange o escopo de organizações, sendo uma recomendação W3C. Entretanto, nem todos os conceitos necessários no escopo do problema encontraram mapeamento único na *ORG Ontology*. Essa limitação gerou sobrecarga semântica em alguns conceitos mapeados (GUIZZARDI, 2005). Por exemplo, Cargo e Atividade, apesar de possuírem significados distintos, onde o primeiro representa um conjunto de atribuições inerentes ao agente público e o segundo um conjunto de atribuições inerentes ao exercício de funções especiais, como chefia e assessoramento, foram mapeados ao conceito **Cargo** (*Post*) da *ORG Ontology*.

A *ORG Ontology* permite ser estendida ou especializada para uso em situações não previstas no padrão, o que resolveria os mapeamentos de conceitos diferentes para uma conceituação igual. Embora isso mitigue o problema da falta de expressividade semântica,

ISBN 978-85-61115-09-8

afeta o benefício descrito na seção 4.4.4. Como o escopo do trabalho era realizar o estudo de caso com tecnologias existentes, não foi realizada a extensão.

4.4 Benefícios do Mapeamento

O mapeamento dos dados seguindo as etapas descritas na seção 2 traz benefícios ao uso dos dados, tanto neste trabalho como em outros.Esses benefícios incluem a: (i) melhor integração dos dados e qualidade de consultas; (ii) possibilidade de inferência; (iii) semântica bem definida e; (iv) utilização de conceitos conhecidos. As subseções a seguir descrevem esses benefícios.

4.4.1 Integração dos dados e qualidade de Consultas.

A realização do mapeamento apresentou melhorias à utilização dos dados. A integração é o principal deles. Com a integração, os dados passaram a ser unificados. A informação se torna mais qualificada dado o maior relacionamento e agregação entre os dados. Também aumenta o relacionamento entre informações diversas e se torna mais fácil o cruzamento de informações, aumentando o poder de análise e melhorando a tomada de decisões baseada nas informações recuperadas.

Isso advém da maior quantidade de dados possíveis de serem utilizados para uma mesma consulta, o que abre a possibilidade de realizar consultas que antes não eram possíveis.

4.4.2 Possibilidade de Inferência nos dados.

O raciocínio automático através de regras de inferência permite que agentes de software entendam e tirem conclusões lógicas sobre os dados existentes. É possível descobrir relações entre dados que se conectam, diretamente ou indiretamente, assim como entre dados que não se conectam, permitindo buscar conexões que não são simples de serem observadas. Por exemplo, no trabalho efetuado é possível através de inferência saber o gasto com pessoal em uma determinada organização, o que não era diretamente possível com a forma original dos dados. Também é possível utilizar inferências para verificar os dados, descobrindo possíveis inconsistências entre eles. Como exemplo, se o usuário possuir dados de orçamentos dos órgãos poderia saber a proporção dos gasto do pessoal em relação ao orçamento do órgão, assim como verificar sua coerência em caso de proporções notadamente inconsistentes.

4.4.3 Utilização de semântica bem definida.

O mapeamento dos dados para uma ontologia permite que a semântica dos dados seja provida pelos conceitos aos quais eles se mapeiam na ontologia. O uso de ontologias com termos bem definidos faz com que os usuários dos dados compreendam sobre o que os dados tratam e aumenta a possibilidade de interpretação correta dos mesmos. Esse mapeamento também afasta a possibilidade de problemas como a Falsa Concordância (GUARINO, 1998), em que os usuários troquem informações sobre elementos diferentes acreditando se tratar do mesmo elemento e não identifiquem essa divergência, levando a interpretação e, consequente, tomada de decisões incorreta. Nesse trabalho, o mapeamento dos dados foi realizado para a ontologia *ORG Ontology*.

4.4.4. Utilização de conceitos amplamente utilizados

Outro benefício do mapeamento é ter os dados mapeados para termos de ontologias que são amplamente utilizadas. O uso de ontologias bastante utilizadas diminui o tempo

ISBN 978-85-61115-09-8

necessário para aprendizado sobre o que as informações tratam e aumentam o seu potencial de uso, dado que vários desenvolvedores e usuários de ontologias já conhecem os seus significados. Isso também aumenta a probabilidade dos dados do trabalho serem usados e cruzados com outros dados, tanto em aplicações de terceiros como em inclusão de novos dados por outros usuários.

Nesse trabalho as ontologias como FOAF (*Friend of a Friend*), SKOS (*Simple Knowledge Organization System*), *vCard* e*GoodRelation* são exemplos de ontologias que possuem seus termos utilizados em larga escala, inclusive por empresas como Google, Wikipedia e IBM.

Conclusão

Dados governamentais são frequentemente publicados de forma não estruturada ou em formatos não padronizados. Frequentemente, diferentes bases de dados utilizam formatos distintos e há pouca ou nenhuma integração entre elas.Neste trabalho foi realizado um estudo de caso utilizando bases de dados governamentais e mapeando os diversos dados para uma ontologia de organizações, denominada *ORG Ontology*, com a publicação de dados integrados baseados nos princípios de Dados Ligados e semanticamente definidos pelos termos de uma ontologia.

Foi realizado um mapeamento conceitual de bases de dados governamentais para a *ORG Ontology*, utilizada como ontologia de referência. Com o uso da linguagem R2RML, o mapeamento conceitual foi codificado e executado, gerando um grafo com mais de 9,4 milhões de triplas,interligando todas as bases de dados entre si e com os conceitos da ontologia. Finalmente, uma aplicação local foi desenvolvida para demonstrar as potencialidades do mapeamento realizado e simular o uso dos dados publicados na Web. Nela é possível realizar consultas específicas com base nos conceitos da ontologia ou em dados individuais (com resultados em diversos formatos), inclusive acessando-os individualmente por meio de seus URIs ou fazendo inferências sobre eles. Ademais, também é possível realizar inferências entre os dados utilizando relacionamentos originalmente somente providos entre os conceitos da ontologia para a qual os dados foram mapeados.

O trabalho mostrou diversos benefícios do uso de uma abordagem com a utilização de dados integrados e semanticamente definidos, conforme descritos na seção 4.4. Em especial, ressalta-se: (i) a integração dos dados, possibilitando maior qualidade na resposta às consultas e maior quantidade de consultas possíveis; (ii) a possibilidade de inferência nos dados, descobrindo-se relações não explicitamente presente nos dados isolados; (iii) a utilização de semântica bem definida, via o mapeamento para uma ontologia de referência, o que aumenta a compreensão dos dados, assim como gera a possibilidade de uso e inclusão de dados por terceiros sem inconsistências (assumindo que se respeite a ontologia) e; (iv) a utilização de conceitos amplamente utilizados pela comunidade diminuindo a curva de aprendizado para desenvolvimento de aplicações por terceiros, facilitando o consumo e análise.

O mapeamento realizado também objetivou a verificação na prática da ocorrência de problemas estruturais ou semânticos, assim como objetivou a verificação dos benefícios para o uso de dados abertos governamentais. Durante a abordagem, foi necessário lidar com problemas, em especial: (i) a falta de identificação única para os elementos recuperados das bases de dados, tanto dentro de uma mesma base de dados quanto entre diferentes bases de dados; (ii) a falta de integração entre bases de dados governamentais, forçando o consumidor a recorrer a suposições para a integração os dados; (iii) a falta de descrição da semântica dos dados publicados, sendo o usuário dos dados responsável por supor a sua semântica com base em rótulos, e; (iv) à falta de expressividade da recomendação W3C *ORG Ontology* para o

ISBN 978-85-61115-09-8

mapeamento de todos os conceitos assumidamente presentes nas bases de dados, levando a sobrecarga semântica de alguns conceitos.

Os problemas encontrados levam a menor confiabilidade das informações. A precisão do grafo gerado pelo mapeamento foi comprometida devido às limitações nas bases de dados, reduzindo a possibilidade de uso das mesmas pela população. Limita-se a possibilidade de tomada de decisões e o correto acompanhamento das políticas públicas utilizando esses dados, uma motivação para a publicação da lei e da disponibilização dos dados em questão.

No trabalho, foram propostas e aplicadas soluções para remediar imediatamente os problemas identificados, de forma a possibilitar a conclusão do trabalho, porém também foram propostas soluções permanentes, a serem tomadas no âmbito governamental para evitar os problemas identificados. Em especial, propõe-se no trabalho a integração das bases de dados governamentais e o seu mapeamento para ontologias de referência, para prover semântica e ajudar a compreensão dos mesmos.

Como perspectivas futuras, maior esforço é necessário para aumentar o escopo do mapeamento de dados para incluir outros elementos, além de servidores e organizações (como gastos e convênios, por exemplo).Isto permitiria gerar uma "nuvem" maior de informações governamentais estruturadas. Essa nuvem de informações possibilitaria um compartilhamento e entendimento maior dos dados governamentais. Também viabilizaria que o acessoàs informações se torne cada vez maior através de aplicações externas ligadas aos dados governamentais. Para o governo de um país que prega transparência da informação, ter os dados publicados de forma como a proposta neste trabalho possibilitaria à população extrair a informação desejada e analisá-la com maior facilidade.

Agradecimentos

Este trabalho contou com o apoio do W3C Brasil, da CAPES, da FAPES (projeto de no. 59971509/12) edo CNPq (projetos de nos. 310634/2011-3 e 485368/2013-7).

Referências

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

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data - The Story So Far. **International journal on semantic web and information systems**, v. 5, n. 3, p. 1–22, 2009.

BORST, W. N. **Construction of Engineering Ontologies for Knowledge Sharing and Reuse**. [s.l.] Universiteit Twente, 1997.

DAS, S.; SUNDARA, S.; CYGANIAK, R. **R2RML: RDB to RDF Mapping Language**. Disponível em: <http://www.w3.org/TR/r2rml/>. Acesso em: 29 abr. 2014.

GUARINO, N. **Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy**. Amsterdam: IOS Press, 1998. p. 3–15

GUIZZARDI, G. Theoretical foundations and engineering tools for building ontologies as reference conceptual models. **Semantic Web**, v. 1, n. 1-2, p. 3–10, 2010.

MCGUINNESS, D. L.; HARMELEN, F. VAN. **OWL Web Ontology Language Overview**. Disponível em: <http://www.w3.org/TR/owl-features/>. Acesso em: 7 maio. 2014.

ISBN 978-85-61115-09-8

REYNOLDS, D. **The Organization Ontology**. Disponível em: <http://www.w3.org/TR/vocab-org/>. Acesso em: 29 abr. 2014.

ISBN 978-85-61115-09-8

UnB-LOD, A Visual Tool to Work With Linked Open Data

Marcus Oliveira Silva
marcus.oli.silva@gmail.com

Rommel Novaes Carvalho
rommel.carvalho@cgu.gov.br

Marcelo Ladeira
mladeira@unb.br

Henrique A. da Rocha
henrique.rocha@cgu.gov.br

Gilson Libório Mende
liborio@cgu.gov.br

Resumo
Integrar dados heterogêneos disponíveis na Web não é uma tarefa trivial porque, em geral, os desenvolvedores devem codificar as transformações e o mapeamento a fim de obter dados integrados. Também não é incomum que usuários com conhecimento básico de TI e dos formatos de representação de dados estejam interessados em utilizar dados abertos em diferentes tarefas, por exemplo, escrever uma reportagem sobre gastos governamentais. Sem uma ferramenta de fácil utilização, os usuários não seriam capazes de tirar partido de todos os dados disponíveis, por exemplo, para utilizar na reportagem todos os dados sobre gastos governamentais disponíveis no portal www.transparencia.gov.br. Esse artigo apresenta o UnB-LOD, uma ferramenta de código aberto, em fase de protótipo, visual e fácil de usar, que permite a usuários com conhecimento mínimo sobre estruturas de dados: (1) transformar conjuntos de dados para conjunto de dados estruturados como RDF; e (2) a fusão de diferentes conjuntos de dados em um conjunto de dados consistente, seguindo os princípios e padrões de dados ligados por meio de modelagem visual e manipulação gráfica de dados.

Palavras-chave: Dados abertos, CSV, Integração de dados, RDF, GUI

Abstract
Integrating the heterogeneous data available in the web is not trivial. Developers usually need to hard code the transformation and the mapping of the data in order to make them useful. Moreover, it is not uncommon to have users, with basic IT or data knowledge, interested in consuming Open Data for different tasks (*e.g.*, to write an article about Government expenditure). Without an easy to use set of tools, users would not be able to leverage on all the data available (*e.g.*, all the Brazilian Government expenditure data available at www.trasparencia.gov.br). This paper introduces UnB-LOD, a visual and easy to use open source tool, still in prototyping phase, that allows users with a minimum knowledge on data structures to perform: (1) the transformation of data sets to RDF structured data sets; and (2) visual data modeling and graphical manipulations to fuse different data sources into a new and consistent one, following the Linked Data principles and patterns.

Key Words: Open Data, CSV, Data Integration, RDF, GUI

Introduction

ISBN 978-85-61115-09-8

The Open Data movement grows rapidly, countries around the world have been publishing information produced by their public bodies on the Web according with specific standards, such as Linked Data (BREITMAN et al., 2012). In what follows will be introduced the UnB-LOD a tool that use Linked Data technology to work with Open Data.

This work has the main focus of supporting the data fusion activities of the Brazilian Office of the Comptroller General (CGU). CGU is the agency of the Federal Government in charge of assisting the President of the Republic in matters which, within the Executive Branch, are related to defending public assets and enhancing management transparency through internal control activities, public audits, corrective and disciplinary measures, corruption prevention and combat, and coordinating ombudsman's activities. CGU has a main branch located in the city of Brasília, where most of the administrative and Information technology (IT) activities are concentrated, and several others local branches, one in each Brazilian capital. Therefore, the vast majority of the CGU's data is stored and maintained in the main branch.

Having most of the IT experts and data concentrated in the main branch has its advantages. However, the problem is how to make this data easily available to the local branches. Not only for consulting, but also for generating new knowledge when fusing with local databases, usually obtained during local auditing.

While the IT professionals have the necessary professional tools and expertise for activities related to Extract, Transform and Load (ETL) and Data Warehouse (DW) processes, they are not able to keep up with the demand of all local branches. The major problem is the lack of those IT professionals in the local branches. However, hiring more IT professionals is not an alternative. Therefore, there is a crucial need from the workers with a minimum IT knowledge to be able to consume and integrate the data. For example, during an investigation, an auditor might want to cross Internal Revenue Service data from the main branch with local public bidding data to find irregularities. Despite the main branch efforts of integrating all the different databases it acquires, it is not possible to forecast all the particularities and specific needs from all investigations in the local branches.

Due to these reasons, the ideal scene is to make the main branch data available to the local branches users and give them an easy to use tool that allow the data integration and manipulation. The following example illustrates the observations and problems presented. In the main branch server there is a database with all the federal employees, their salary, etc. However, there is no data about the local government employees. In the local branch, an auditor might have that data available or acquire it during an investigation or with a partnership with the local government. This data could be in any known format, e.g., CSV dumps or SQL endpoints, etc. In this scenario, he/she may want to combine these data in order to check if there is an accumulation of illegal public functions (in Brazil, only professors and health professionals can have more than one public job).

Section 1 introduces the UnB-LOD. And Section 2 explains the architecture of UnB-LOD. An study case is detailed in the Section 3, showing different approaches to solve Open Government Data (OGD) integration problems, one of these approaches is the use of the UnB-LOD features. Section 4 shows related work and Section 5 has the conclusion.

1 The UnB-LOD Proposal

UnB-LOD is a free, open source, and easy to use tool for loading and integrating data. Although its use has been focused on government data, UnB-LOD can be used to manipulate any kind of Open Data (ACCAR, ALONSO, NOVAK, 2009). UnB-LOD adopts Linked Data (BIZER, HEATH, BERNERS-LEE, 2009) standards. One of the reasons for using Linked Data is that Brazil is one of the co-founders of the Open Government Partnership

ISBN 978-85-61115-09-8

(TAUBERER, 2014), and it is committed to public transparency and the publication of official data (BREITMAN et al., 2012). Besides, as pointed by the W3C:

The desire for an open and transparent government is more than open interaction and participation, appropriate data as products of the government must be shared, discoverable, accessible, and able to be manipulated by those desiring the data. The data as well must be linked via subject, relevance, semantics, context, and more. Linked Data offers the information consumer ways and means to find relevant and pertinent information through search, queries, interfaces, or tools available today and for tomorrow (ACCAR, ALONSO, NOVAK, 2009).

UnB-LOD has a visual appeal, the user will do all the work through a graphical user interface (GUI). The UnB-LOD builds on top of other tools and frameworks, which will be described later. Among the Linked Data standards followed by UnB-LOD are the RDF[4] (Resource Description Framework) and URI[5] (Uniform Resource Identifier) to represent and to manipulate the data (HELBER et al., 2009). With UnB-LOD a user can import data from different sources and of different types (e.g., CSV Brazilian Government expenditure data), which will then be transformed to RDF data associated to a simple RDFS[6] (RDF Schema) ontology to describe the meta-data. Having imported and structured the data it will be possible to integrate and fuse the data.

The following list describes the pipeline of the UnB-LOD's work-flow:

1. Preparation
   a) Model and write the dataset
2. Data Collection
   a) Write the data importation schema
   b) Import data from different sources
   c) Transform data to structured RDF
3. Data Integration
   a) Multiple data mapping and data fusion
   b) Output a new and consistent data set

UnB-LOD is implemented in Java and it is modularized using Eclipse Rich Client Application7 (RCP), eclipse RCP provides an easy way to create desktop applications with industry standards. UnB-LOD is still in the prototyping stage.

2 UnB-LOD architecture

UnB-LOD is divided in three main phases, preparation, collection and data integration. Figure 1 illustrates the UnB-LOD architecture and Figure 2 shows the UnB-LOD user interface.

4 http://www.w3.org/RDF/
5 http://www.w3.org/wiki/URI
6 http://www.w3.org/TR/rdf-schema/
7 http://www.eclipse.org/home/categories/rcp.php

ISBN 978-85-61115-09-8

Figure 1 – UnB-LOD architecture

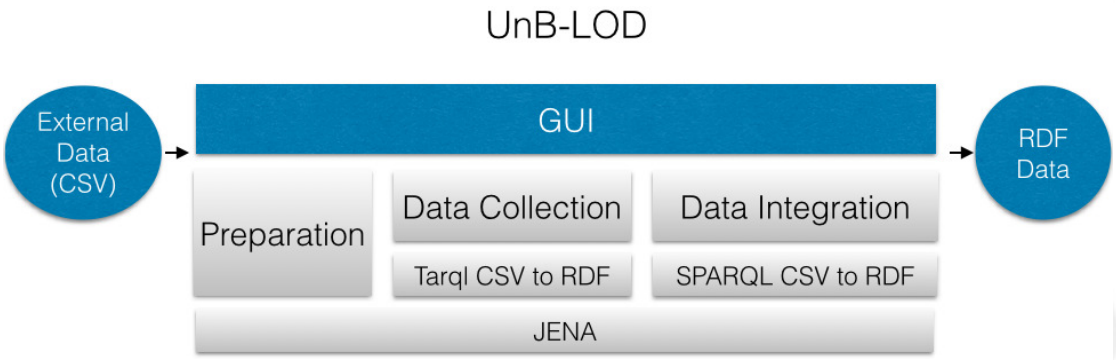
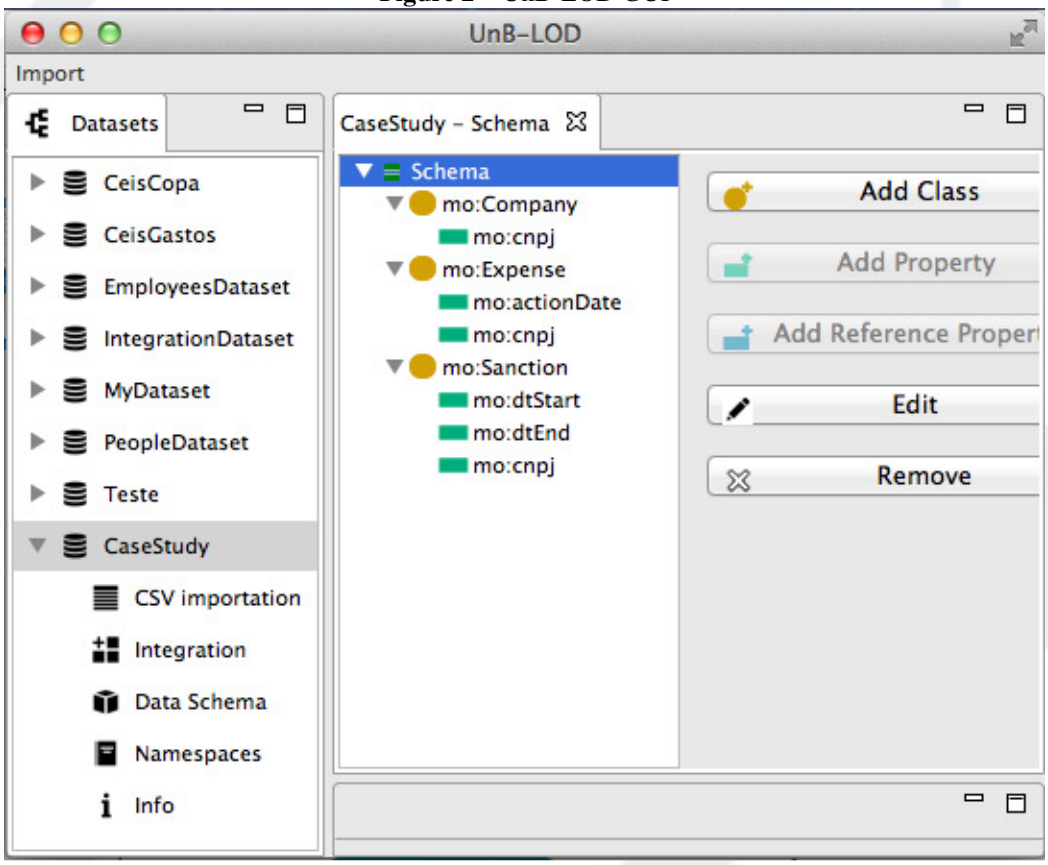


Figure 2 – UnB-LOD GUI



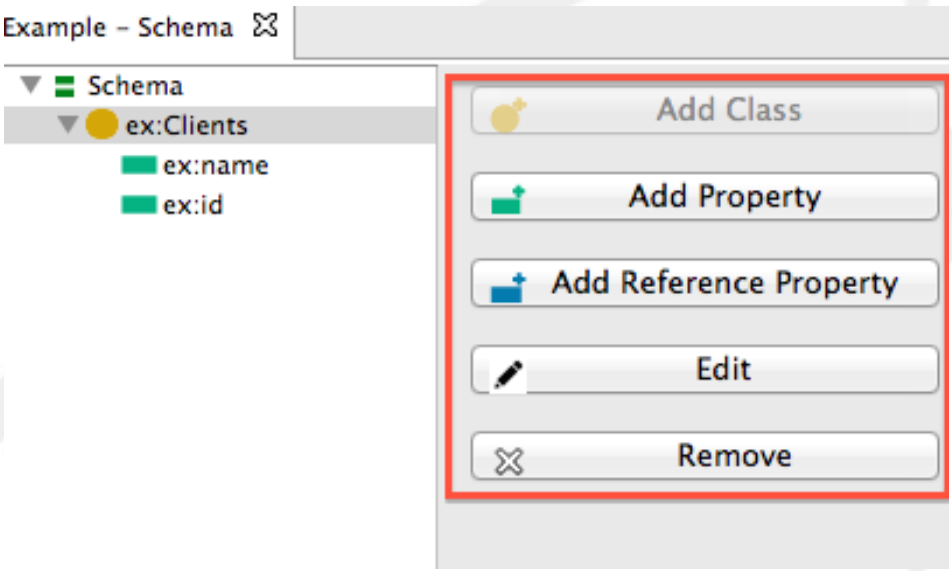
Preparation.To perform data integration using UnB-LOD it is necessary to model a data schema. Each dataset has an associated schema. This schema is similar to RDFS[8], but simpler. It only has the *Class*, *Property* and *ReferenceProperty* components. The schema was simplified because these components are the minimum required to have an expressive data schema. Future versions will include more components of the RDFS, like inheritance and

8 http://www.w3.org/TR/rdf-schema/

ISBN 978-85-61115-09-8

collections. Using UnB-LOD GUI it is possible to define classes, properties, and relationship between classes. Figure 3 shows the GUI to build a data schema. In the left there is tree representing the schema, there is a class named *ex:Client* with properties *ex:name* and *ex:id*, in the right there are the buttons to add more properties and classes to the schema.

Figure 3 – UnB-LOD Schema GUI



Data Collection. In this phase the user will map and transform, using the GUI, each data entity to generate a structured RDF dataset from the CSV input only. The data collection phase works on the Tarql API9, which is a tool to map CSV to RDF using SPARQL syntax.

Listing 1 – Code 1

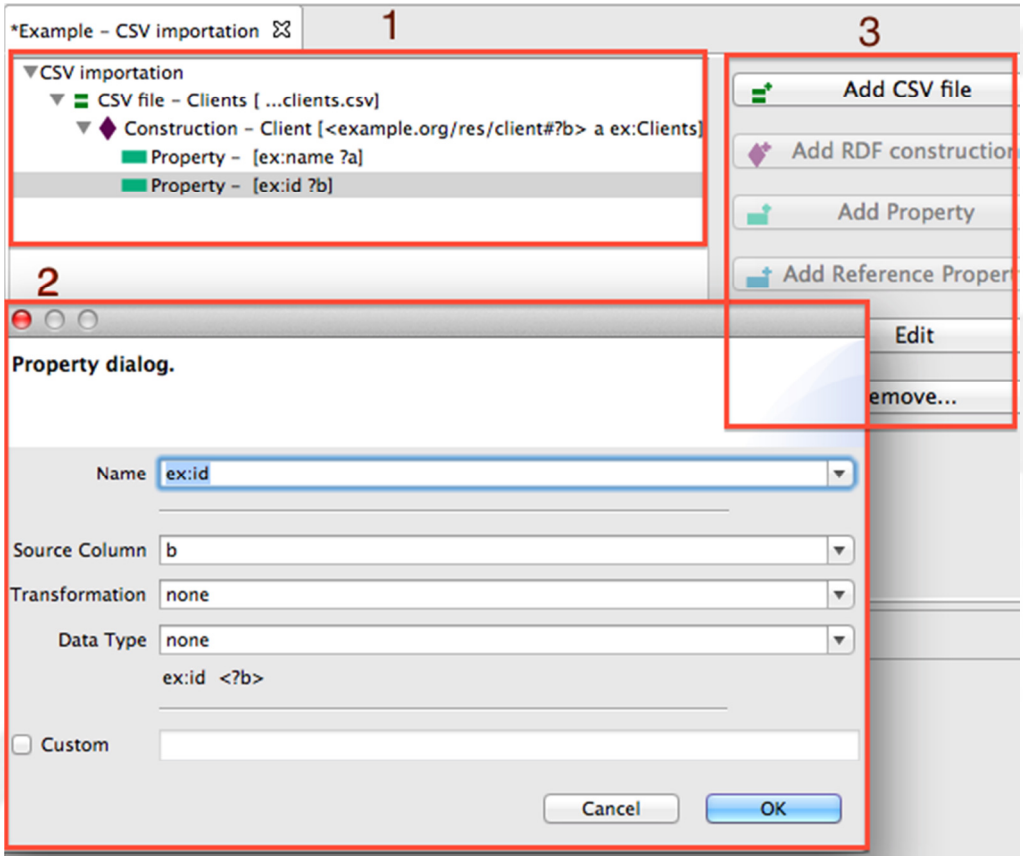
```
1  CONSTRUCT {
2    ?URI a ex:Clients;
3         ex:name ?a;
4         ex:id ?b;
5  }
6  FROM <file:clients.csv>
7  WHERE {
8    BIND (URI(CONCAT('clients#', ?b)) AS ?URI)
9  }
10 OFFSET 1
```

The Listing 1 shows a CSV to RDF mapping using tarql. A CSV file of clients is mapped to a RDF graph. The line 8 shows how to build a URI of the RDF resource based on column B. Columns A and B of the CSV are mapped to the properties *ex:name* and *ex:id*. Figure 4 shows the GUI to build a CSV to RDF mapping. Block 1 has a tree that is visual representation of the mapping. Block 2 shows a dialog to input/edit values of a property/column mapping. This dialog is opened when *Add Property* button or *Edit* button are clicked.

9 https://github.com/cygri/tarql/

ISBN 978-85-61115-09-8

Figure 4 – UnB-LOD CSV to RDF mapping GUI



Data Integration. In this phase the user will use the newly structured RDF datasets and will define the vocabulary mappings and data fusion using UnB-LOD's GUI. The data integration phase works on top of the Jena Framework[10], UnB-LOD uses the *CONSTRUCT* syntax from the Jena SPARQL query engine to generate new integrated RDF data.

3 Case study

The objective of this case study is to verify if companies are appearing in direct government expenses data during its period of sanction or suspension, which would be an indicator of a possible illegal action, since companies sanctioned or suspended theoretically could not be receiving money from governmental institutions.

This case study was aimed to conduct an investigation upon the integration of the following databases:

1) CEIS[11]: this database is the national register of sanctioned/suspended companies. These companies cannot celebrate new contracts with the Government during the time they are suspended;

10 https://jena.apache.org/
11 http://www.portaldatransparencia.gov.br/ceis/Consulta.seam

ISBN 978-85-61115-09-8

2) Direct expenditures[12]: this database has information on all Federal Government direct expenditures since 2010. This information is updated daily.

Table I shows the CSV header from the CEIS dataset and Table II shows the CSV header from the government direct expenditures dataset.

Table 1 – CSV colunms names of CEIS data

Column	**Name**
A	StartSanction_Date*
B	EndSanction_Date*
C	agency
D	agency_location
E	info_origin
F	Date_Info_Origin
G	Sanction_Type
h	Company_Name
i	Company_Location
j	Company_CNPJCPF*

Table 2 – CSV colunms names of government direct expenditure data

Column	**Name**
A	CNPJ*
B	Company_Name
C	Fantasy_Company_Name
D	CNAE_Cod
E	Company_Type_Cod

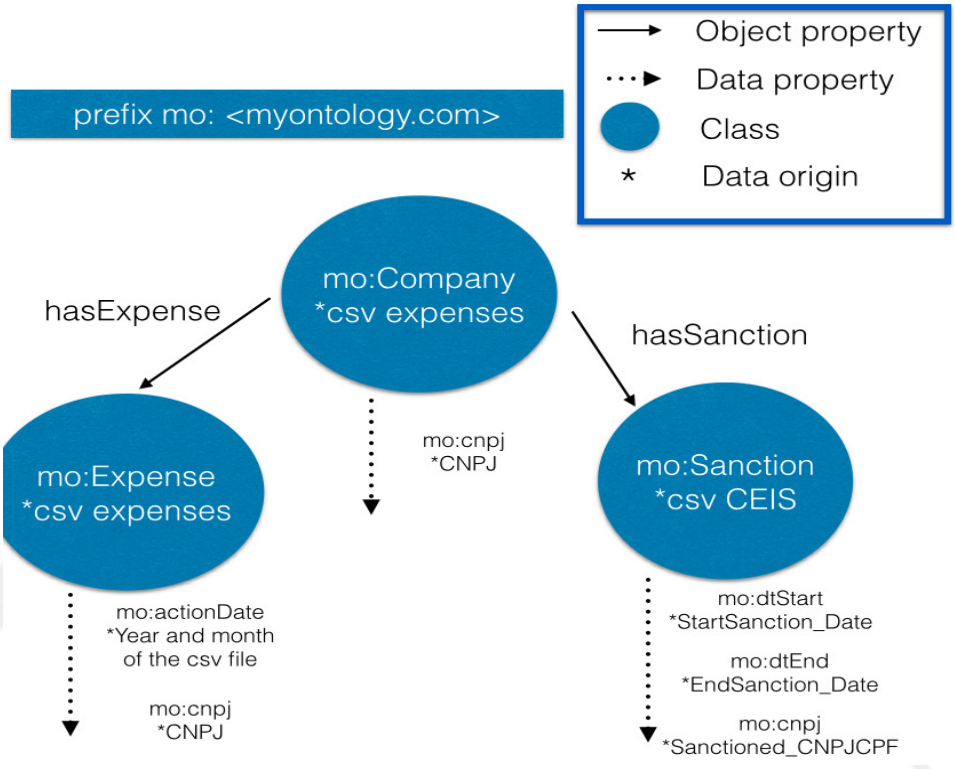
The names of the columns are in Portuguese in the original file, a translation was made for better understanding of the text. The columns names followed by * are the ones used in this case study. The CNPJ is the company ID. The objective is to check if a CNPJ in the CEIS dataset match a CNPJ in of the government direct expenditures, during the period defined by the dates in the columns *StartSanction_Date* and *EndSanction_Date*.

Figure 5 shows a dataset schema which represents the fusion of the CEIS dataset and the direct expenses data. This figure explains the origin of each value.

12 http://www.portaltransparencia.gov.br/PortalComprasDiretasPrincipal2.asp

ISBN 978-85-61115-09-8

Figure 5 – Schema CEIS/Expenses



Manual approach: To accomplish this data integration example, with a common and manual approach, a basic user needs to: download all data, use some spreadsheet editing tool, making CSV joins manually, and make filters available in spreadsheets software (e.g. Excel, Google Refine). Besides, he/she needs to do all the work without a methodology or work-flow to guide the process. This approach requires a lot of manual work, it is subject to many errors and has no reuse mindset. Nevertheless, a more efficient and intelligent approach requires more knowledge and skills.

Next, an approach using SPARQL mapping codes and tools to perform the data integration in a more automated and efficient way will be presented. Moreover, it will also be shown how this task can be significantly facilitated when using the UnB-LOD GUI, making it accessible to more users.

Automated approach: In order to do the integration, a user needs to collect and transform the data to a new representation that is useful to his/her purpose. He/She can do that using some tool or coding some script, which is hard for users with little background on data integration.

The Tarql mapping (Listing 2) shows how to perform the transformation of a CSV dataset into an RDF dataset following the integration schema suggested by Figure 5. Lines 5, 18, 47, and 64 shows a *CONSTRUCT* query that generates RDF resources to the *mo:Company*, *mo:Sanction* and *mo:Expense* classes. Each row in the CSV files, specified by the *FROM* expression, is mapped to an RDF resource. All the columns of the CSV files can be used as variables in the *CONSTRUCT* query (e.g., *?CNPJ* and *?sanctionStartDate*). The *BIND* keyword binds a variable to an expressions that uses values from columns, the special variable *?URI* is created by a bind expression and represents the URI of the resource.

ISBN 978-85-61115-09-8

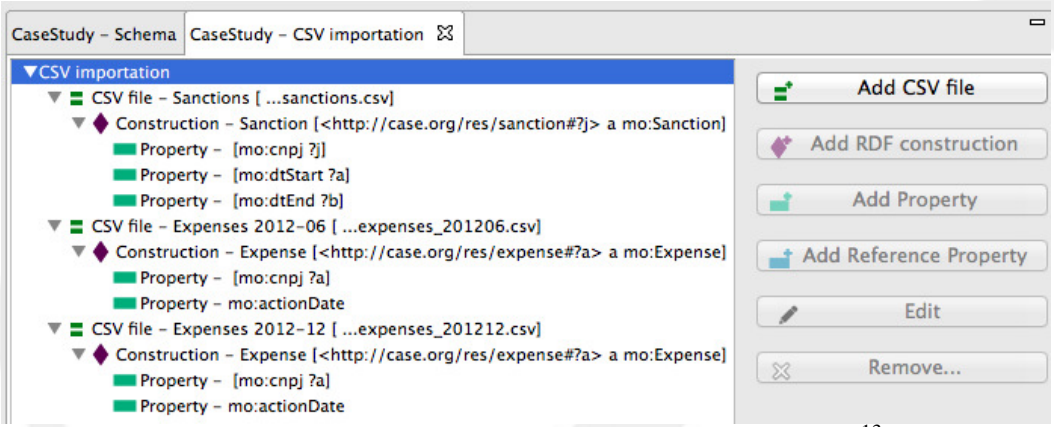
Listing 2 – Code 2

```
PREFIX mo: <http://myontology.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

#Company
CONSTRUCT {
  ?URI a mo:Company;
       mo:cnpj     ?CNPJ;
}
FROM <file:data/expenses.csv>
WHERE {

  BIND ('http://case.org/res/company#' AS ?URI_PREFIX)

  BIND (URI(CONCAT(?URI_PREFIX, ?CNPJ)) AS ?URI)
}

#Sanction
CONSTRUCT {
  ?URI a mo:Sanction;
       mo:cnpj          ?cnpj;
       mo:dtStart       ?sanctionStartDate;
       mo:dtEnd         ?endSanctionDate;
}
FROM <file:data/sanctions.csv>
WHERE {
  BIND ('http://case.org/res/sanction#'  AS ?URI_PREFIX)

  BIND (URI(CONCAT(?URI_PREFIX, STRUUID())) AS ?URI)

  BIND ( (replace(?Company_CNPJCPF, "\\.", "")) AS ?v1)

  BIND ( ?StartSanction_Date AS ?d1)
  BIND ( ?EndSanction_Date AS ?d2)

  BIND ( strdt(concat(substr(?d1, 7, 4), '-', substr(?d1, 4, 2) , '-',
         substr(?d1, 1, 2), 'T00:00:00'), xsd:dateTime)
    AS ?sanctionStartDate)

  BIND ( strdt(concat (substr(?d2, 7, 4), '-', substr(?d2, 4, 2) , '-',
         substr(?d2, 1, 2), 'T00:00:00'), xsd:dateTime)
   AS ?endSanctionDate)

}
OFFSET 1
```

```
#Expenses jun/2012
CONSTRUCT {
  ?URI a mo:Expense;
       mo:cnpj       ?CNPJ;
       mo:actionDate ?actionDate;
}
FROM <file:data/expenses_201206.csv>
WHERE {
  BIND ('http://case.org/res/expense#'   AS ?URI_PREFIX)

  BIND (URI(CONCAT(?URI_PREFIX, STRUUID())) AS ?URI)

  BIND ( strdt(concat( str("2012"),       '-', str("06") , '-', '
        01', 'T00:00:00'), xsd:dateTime) AS ?actionDate)
}
OFFSET 1

#Expenses nov/2012
CONSTRUCT {
  ?URI a mo:Expense;
       mo:cnpj    ?CNPJ;
       mo:actionDate  ?actionDate;
}
FROM <file:data/expenses_201212.csv>
WHERE {
  BIND ('http://case.org/res/expense#'  AS ?URI_PREFIX)

  BIND (URI(CONCAT(?URI_PREFIX, STRUUID())) AS ?URI)

  BIND ( strdt(concat( str("2012"), '-', str("12") , '-',
        '01', 'T00:00:00'), xsd:dateTime) AS ?actionDate)

}
OFFSET 1
```

UnB-LOD makes the task of importing data more intuitive and easier. Figure 6 shows the equivalent mapping from the code above, using UnB-LOD's GUI, a tree modeling representation. The "add CSV File" button in the figure is equivalent to writing a SPARQL *FROM* expression, the *add RDF Construction* is equivalent to writing a SPARQL *CONSTRUCT* query, and *add Property* or *add Reference Property* is equivalent to writing *BIND* expressions and using binded variables as values of the RDF properties in the CONTRUCT block.

Figure 6 – UnB-LOD CSV to RDF schema



This import would generate some RDF data (show in turtle [13]) in the Listing 3:

13 http://www.w3.org/TR/turtle/

ISBN 978-85-61115-09-8

Listing 3 – Code 3

```
@prefix mo: <http://myontology.org/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

#(1)
<http://case.org/res/expense#557c7460-7c90-44a6-b978-14f06731fe4d>
     a        mo:Expenses ;
     mo:cnpj "64926041000196" ;
     mo:actionDate "2012-12-01T00:00:00"^^xsd:dateTime .

<http://case.org/res/expense#5565ac1b-8af1-4e13-a087-cdc43fd348bf>
     a        mo:Expense ;
     mo:cnpj "14961882000166" ;
     mo:actionDate "2012-12-01T00:00:00"^^xsd:dateTime .

#(2)
<http://case.org/res/company/64926041000196>
       a        local:Company ;
       mo:cnpj          64926041000196;
...
```

With the output code in Listing 3, since the data is now fused into a single and consistent dataset representation, the user can now perform SPARQL queries to find, for example, the sanctioned/suspended companies which were earning money from the Government in a specific time frame. Applying a SPARQL query in this integrated data it was possible to discover that 219 companies from 44579 were earning money during a sanctioned/suspended time. There isn't a GUI to write SPARQL queries easily in the UnB-LOD yet, but it is planned to include a graphic SPARQL query builder in the future.

4 Related Work

There is an enterprise software, called Topbraid Composer, that has a semantic data integration visual environment. Topbraid Composer is a modeling tool and a collection of integrated Semantic Web solutions. But Topbraid Composer is a proprietary software and has payed licenses. Topbraid Composer is used by Semantic Web experts while UnB-LOD focuses on non-experts users, having a simple interface and more intuitive concepts to work with Linked Open Data. For more details, see http://www.topquadrant.com/tools/ide-topbraid-composer-maestro-edition/.

There are several free tools that transforms or extract RDF from spreadsheets and CSV. NOR2O is a library for transforming non-ontological resources to ontologies. The last version was released in 2010-11-26. Now a days, there is no technical support for this library. Apache Anything To Triples (Any23) is a library, a web service and a command line tool that extracts structured data in RDF format from a variety of Web documents, including CSV. RDF123 is an application and web service for converting data in simple spreadsheets to an RDF graph. RDF123 was released in 2008 at University of Maryland Baltimore County. The required background on IT to use these tools are higher than the one required to use UnB-LOD. UnB-LOD is ease to use for no IT expert because the its GUI. The UnB-LOD's user need have no deep knowledge in IT because the CSV to RDF schema is guided by buttons (as show in Figure 6).

ISBN 978-85-61115-09-8

5 Conclusion

Data collection and integration can be a challenging task. On the one hand, a manual approach requires more work, it is subject to errors, and it usually cannot be reused once new data becomes available. On the other hand, using more advanced techniques, like writing some scripts or modeling data mappings requires specific knowledge and skills. This paper presents UnB-LOD, a visual tool that can simplify the work, increase the reuse, and decrease the required knowledge, while still allowing more complex data manipulations to be performed. These manipulations can also be accomplished using tools like spreadsheets. However, it requires much more work than using UnB-LOD.

Future work will allow: exporting the data to different formats (e.g., XML, SQL); the use of inheritance in the schema modeling; scheduling tasks to make the UnB-LOD more automated; and a GUI to build SPARQL queries.

References

ACCAR, S.; ALONSO, J.; NOVAK , K. Improving Access to Government through Better Use of the Web. W3C Interest Group, 2009.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), v.5, n.3, pp. 1-22, doi:10.4018/jswis.2009081901, 2009.

BREITMAN, K. S.; CASANOVA, P.; SARAIVA, M.A.; VITERBO, D.; MAGALHÃES, J.; FRANZOSI, R.P.; CHAVES, E. Open Government Data in Brazil. Intelligent Systems, IEEE. v. 27:3, pp. 45-49, 2012.

HELBER, J.; FISHER M.; BLACE, R.; PEREZ-LOPEZ, A. Semantic Web Programming. Indianapolis: Wiley, 2009.

TAUBERER, J. Open Government Data: The Book. 2sn ed. Civic Impulse LLC, http://opengovdata.io/, 2014

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

GERAÇÃO SEMIAUTOMÁTICA DE ITENS A PARTIR DE DADOS ABERTOS PARA AVALIAÇÕES EDUCACIONAIS COM O USO DE TESTES ADAPTATIVOS COMPUTADORIZADOS

Paulo Rogério Pires Manseira
paulo.manseira@sociesc.org.br
Mehran Misaghi
mehran@sociesc.org.br

Resumo
Este trabalho apresenta uma proposta para geração semiautomática de itens para avaliações educacionais de baixo risco. O objetivo do estudo é apresentar uma forma para a Geração Automática de Itens (GAI) a partir da base de dados abertos DBpedia, de forma que os itens gerados possam ser incorporados a um Banco de Itens (BI) possibilitando ampliar o número de oportunidades para o uso de Testes Adaptativos Computadorizados (TACs). A abordagem metodológica é composta por pesquisa bibliográfica para a identificação de conceitos e tecnologias necessárias para sua posterior aplicação em um estudo de caso onde se verifica a viabilidade da GAI a partir da bases de dados abertos DBpedia. Percebe-se a viabilidade prática da proposta, bem como identificam-se os limites do trabalho e seus encaminhamentos futuros.

Palavras-chave: Geração Semiautomática de Itens. Testes Adaptativos Computadorizados. Dados Abertos. DBpedia.

Abstract
This paper presents a proposal for semi-automatic items generation of low risk educational assessment. The aim of the study is to present a way for automatic generation of items (AIG) from the base of open data DBpedia, so that the items generated can be incorporated to anItem Bank enables to expand the number of opportunities for using Computerized Adaptive Testing (CAT). The methodological approach consists of literature search to identify concepts and technologies required for its subsequent application in a case study where it verifies the feasibility of AIG from the base of open data DBpedia. Realizes the practical feasibility of the proposal and identifies the limits of the work and its future directions.

Key Words: Semi-automatic Item Generation. Computerized Adaptive Testing. Open Data. DBpedia.

Introdução

Atualmente as instituições de ensino têm buscado obter e explorar conjuntos de dados acerca de seu público-alvo com o objetivo de diminuir a evasão e proporcionar maior qualidade de seus serviços educacionais ao mesmo tempo que identificam com maior clareza o perfil de seus alunos de forma a personalizar o processo de ensino-aprendizagem (JOHNSON et al., 2013).

ISBN 978-85-61115-09-8

A personalização do processo de ensino-aprendizagem permite o alcance de benefícios como o estímulo à postura ativa do estudante, respeito a seu próprio ritmo de aprendizagem e ênfase na formação com feedback sistemático e contínuo. Logo, para alcançar tais benefícios é necessário o uso constante e regular de testes diagnósticos e formativos que gerem informações para análise da proficiência (nível de conhecimento), e que exige, portanto, maior esforço na elaboração das avaliações, maior tempo na análise dos resultados e na posterior provisão do feedback de cada aluno por parte dos professores (ALVES et al., 2011).

Uma possibilidade para atenuar esse problema é a automatização das atividades de aplicação e análise dos resultados de testes diagnósticos e formativos através de Testes Adaptativos Computadorizados (TACs) baseados na Teoria de Resposta ao Item (TRI) (SCHEUERMANN; BJÖRNSSON, 2009; MOREIRA JUNIOR, 2011). Contudo, um dos principais aspectos que limitam a adoção de TAC é a construção e manutenção do Banco de Itens (BI), devido a grande quantidade de questões que devem ser formuladas de forma a garantir a confiabilidade estatística e o não conhecimento prévio dos itens que serão aplicados em cada teste individualizado para cada examinando, bem como evitar uma seleção inadequada de itens para cada teste e assim prejudicar a estimativa do nível de habilidade do indivíduo (MOREIRA JUNIOR, 2011). Na busca por mitigar essa desvantagem, a Geração Automática de Itens (GAI) vem ganhando interesse há alguns anos, particularmente por promover a diminuição de custos e tempo na autoria de itens a serem utilizados em diversos tipos de testes. Trata-se de uma metodologia que permite a geração de itens para testes a partir de um modelo de item que permite definir as variáveis, o corpo da questão, suas opções e quaisquer outras informações auxiliares (GIERL; LAI, 2013).

Assim, o objetivo deste trabalho é apresentar uma forma para GAI a partir de uma base de dados abertos (Linking Open Data, LOD, em inglês), de forma que os itens gerados possam ser incorporados a um BI.

Espera-se dessa maneira possibilitar a ampliação do número de oportunidades para o uso de TAC, com foco em avaliações de baixo risco, como por exemplo, testes diagnósticos e testes formativos, que não apresentam consequências diretas em relação à medidas e cálculos de resultados acadêmicos.

Com este intuito, a abordagem metodológica deste trabalho é composta por pesquisa bibliográfica para a identificação de conceitos centrais sobre TRI, TAC, GAI e LOD para sua posterior aplicação em um estudo de caso onde se verifica a possibilidade da geração de itens a partir de dados abertos da base DBpedia a serem consumidos futuramente em um TAC.

Portanto, o presente trabalho está organizado em quatro seções. Na primeira seção apresenta-se a introdução do trabalho, objetivos e relevância para o seu desenvolvimento. Na segunda seção, são expostos os fundamentos conceituais da pesquisa. Adiante, na terceira seção é apresentado o estudo de caso onde busca-se identificar a viabilidade prática para GAI com o uso de dados abertos da base DBpedia, trabalho que está em andamento. Na quarta seção são apresentadas as considerações finais juntamente com os resultados esperados e as etapas a serem realizadas.

1 Fundamentos conceituais

O TAC é um teste computadorizado flexível e adaptável, em tempo real, ao indivíduo que está sendo examinado (PITON-GONÇALVES, 2012). Seu surgimento está relacionado aos desdobramentos sobre os estudos de técnicas psicométricas e à evolução das tecnologias de *hardware* e *software* que tornaram possíveis a execução rápida e eficiente dos complexos cálculos utilizados pelo modelo psicométrico da TRI proposto por Frederick Lord e Melvin Novick em 1968 (MOREIRA JUNIOR, 2011). Uma necessidade premente para a

ISBN 978-85-61115-09-8

implementação de TACs é a construção de um BI com grande quantidade de questões, que pode esbarrar na eventual falta de profissionais especialistas em cada assunto a ser avaliado nos testes, ou na falta de tempo hábil para a formulação da quantidade adequada de itens. Assim, explora-se a possibilidade de GAI a partir de modelos de questões com acesso a bases de LOD disponíveis livremente na *Internet*. Portanto apresenta-se na sequência os conceitos de TRI, TAC, GAI e LOD que são aplicados no estudo de caso em andamento que é relatado neste trabalho.

1.1 TRI – Teoria de Resposta ao Item

A TRI é um conjunto de modelos estatísticos que buscam medir algum traço latente do indivíduo (COSTA, 2009; PASQUALI, 2011). É um componente da Psicometria, uma área de estudos das medidas em Psicologia, que tem suas bases na Teoria do Traço Latente (uma característica do indivíduo que não se pode medir diretamente, mas que pode ser inferida a partir de questões que são apresentadas em um teste, por exemplo), advinda da Psicologia, e na Estatística e Probabilidade de onde empresta os instrumentos de medição e descrição (PASQUALI, 2011).

A TRI tem como unidade de análise o item que é administrado em um teste, sugerindo assim uma forma de representar por meio de modelos matemáticos a relação entre 3 variáveis: a medida do traço latente de um indivíduo, as características de determinado item apresentado em um teste e a probabilidade de tal indivíduo dar a resposta certa ao item administrado. Isso é tornado possível pois tanto o item administrado quanto a medida do traço latente do indivíduo são colocados em uma mesma escala (MOREIRA JUNIOR, 2011).

Para que a análise a partir do item seja válida, são considerados dois pressupostos teóricos para o modelo unidimensional: a independência local e a unidimensionalidade. A independência local postula que para um dado traço latente as respostas apresentadas a diferentes itens de um teste são independentes, ou seja, não existe influência entre respostas dadas a outros itens do mesmo teste, o item permite verificar o nível do traço latente do indivíduo que respondeu e consequentemente permite verificar se o item respondido é mais ou menos difícil quando considerado o resultado do teste, sempre em função do traço latente que está sendo medido. A unidimensionalidade postula que há apenas um traço latente responsável, ou dominante, pela resposta a todos os itens apresentados em um teste, e que todos os itens apresentados no teste medem apenas um, e o mesmo, traço latente, ou seja, a estrutura interna do item refere-se a um único assunto e uma única habilidade requerida para sua solução (PASQUALI, 2011).

Entre os diversos modelos encontrados identificou-se que o Modelo Logístico com 3 Parâmetros (ML3) tem sido mais utilizado em situações onde se procura medir o nível de proficiência de um estudante em ambientes educacionais (MOREIRA JUNIOR, 2011). Este modelo leva em consideração para cada item a ser apresentado durante o teste, o seu nível de discriminação, o seu nível de dificuldade e, considerado o nível de habilidade estimado do respondente, qual sua probabilidade de acerto casual (chute). Para uma lista abrangente dos modelos existentes sugere-se uma consulta a Moreira Junior (2011).

A fórmula do modelo ML3 é (PASQUALI, 2011):

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta_j - b_i)}} \quad (1)$$

Onde:

i indica um determinada questão do teste que pode assumir qualquer valor de 1 a *I*;

j indica um determinado indivíduo que realizou o teste e pode assumir qualquer valor de 1 a *n*;

ISBN 978-85-61115-09-8

U_{ij} é uma variável que assume o valor 1 ou 0, caso o indivíduo j acerte ou erre o item i, respectivamente;

θ_j é o traço latente estimado, como por exemplo sua proficiência, do indivíduo j;

a_i é o parâmetro de discriminação do item i;

b_i é o parâmetro de dificuldade do item i;

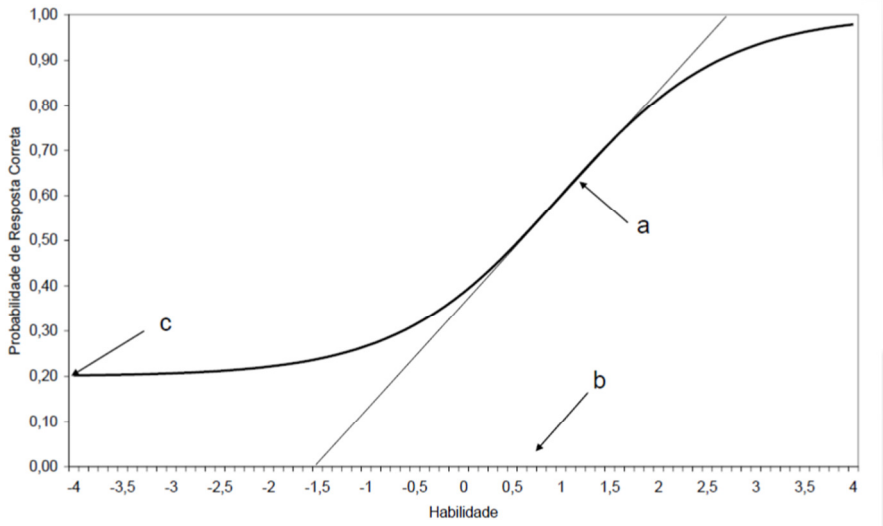
c_i é o parâmetro de acerto casual, que indica a probabilidade de um indivíduo com baixo nível estimado acertar casualmente a questão;

e é a constante matemática Número de Euler, base dos logaritmos naturais, cujo valor é 2,718281...;

$P(U_{ij}=1|\theta_j)$ é a Função de Resposta do Item (FRI) e indica a probabilidade do indivíduo j com traço latente medido em θ_j responder corretamente o item i, ou seja, é a proporção de respostas corretas para o item i entre os indivíduos com o traço latente medido em θ_j.

A representação gráfica da FRI é uma curva em forma de S, representada no Gráfico 1, onde se pode observar a relação entre os parâmetros de dificuldade, discriminação e acerto casual em função da medida do traço latente e da probabilidade de resposta correta ao item, e recebe o nome de Curva Característica do Item (CCI).

Grafico 1 – Curva Característica do Item (CCI)



Fonte: Adaptado de Moreira Junior, 2011

No Gráfico 1 são usados os valores a=1,2, b=0,6 e c=0,2 e uma escala de dificuldade com média zero e desvio padrão 1, pois o que interessa nessa escala é a relação existente entre os diferentes pontos da escala e não necessariamente seus valores, que podem variar de -∞ a +∞ (PASQUALI, 2011). Assim, verifica-se que o eixo Y indica a probabilidade de um indivíduo acertar uma determinada questão enquanto o eixo X representa a dificuldade da questão (parâmetro b, que é medido na mesma escala do traço latente θ) quando o desenho da curva cruza a linha de 50% de probabilidade de acerto.

No exemplo apresentado, o parâmetro c é indicado pela assíntota inferior da curva, no ponto 0,2 da escala, indicando a existência de 20% de probabilidade de acerto casual por indivíduos com traço latente estimado menor que o nível de dificuldade do item, que é representado pelo parâmetro b. O parâmetro a, que indica o nível de discriminação do item, é representado pela inclinação ou ângulo da curva no ponto de inflexão quando a curva cruza a linha de 50% de probabilidade de acerto. Valores baixos no parâmetro a indicam que tanto indivíduos com níveis estimados no traço latente mais altos quanto mais baixos tendem a

ISBN 978-85-61115-09-8

acertar a questão, enquanto que valores mais altos tendem a dividir a população em apenas 2 grupos, aqueles que tem o traço latente medido abaixo do valor de *b* e aqueles que o tem acima do valor de *b*. Assim, não são desejáveis valores muito baixos ou muito altos, pois isso dificulta a distinção entre indivíduos com níveis estimados de traço latente muito próximos uns dos outros. Segundo Moreira Junior (2011) valores no parâmetro de discriminação a partir de 0,7 até valores em torno de 1 são considerados bons para os itens.

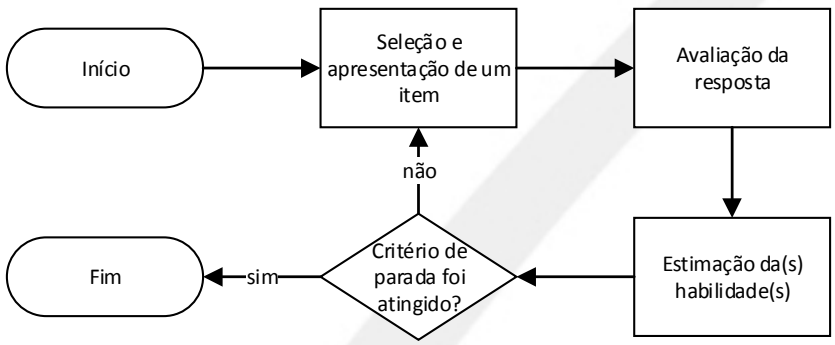
Logo, percebe-se que quanto maior o nível do traço latente de um indivíduo maior é a probabilidade de acertar o item apresentado durante um teste, e quanto menor o nível do traço latente menor será a probabilidade de acerto. Dessa maneira, desde que os itens sejam construídos de maneira adequada e estejam calibrados em uma métrica comum tanto para dificuldade do item quanto para o nível de traço latente, e que seja garantida sua unidimensionalidade, o traço latente do indivíduo apresenta-se invariante, ou seja, ele é independente dos itens utilizados para medi-lo, já que considerando seu nível estimado em determinado tempo *t*, ele não muda, quer o teste aplicado seja fácil ou difícil.

1.2 TAC – Testes Adaptativos Computadorizados

Testes Adaptativos Computadorizados são testes administrados através de um sistema informatizado que apresenta questões e coleta as respostas do indivíduo que está sendo testado, sendo que sua principal característica é a escolha em tempo real dos itens que são administrados ao examinando de acordo com a capacidade do mesmo, que é estimada à medida que responde a cada item, gerando assim um teste personalizado para cada indivíduo (MOREIRA JUNIOR, 2011; PITON-GONÇALVES, 2012).

A partir de um critério determinado o sistema escolhe a primeira questão a ser apresentada e, na sequência, de acordo com sua resposta, correta ou incorreta, a próxima questão é selecionada em um BI, permitindo dessa maneira que o teste seja formado por questões adequadas ao traço latente que está sendo avaliado. O teste segue essa lógica até que se encontre um ponto de equilíbrio, ou qualquer outro critério de parada, na estimativa do nível de proficiência (FETZER et al., 2011), conforme ilustrado na Figura 1. Ao final do teste tem-se uma estimativa bastante precisa do nível de domínio que o indivíduo apresenta sobre o conteúdo avaliado. Diversos estudos demonstram que o TAC precisa de menos itens do que os testes convencionais para atingir seu objetivo (VAN DER LINDEN; GLAS, 2010).

Figura 1 – Algoritmo básico de um TAC



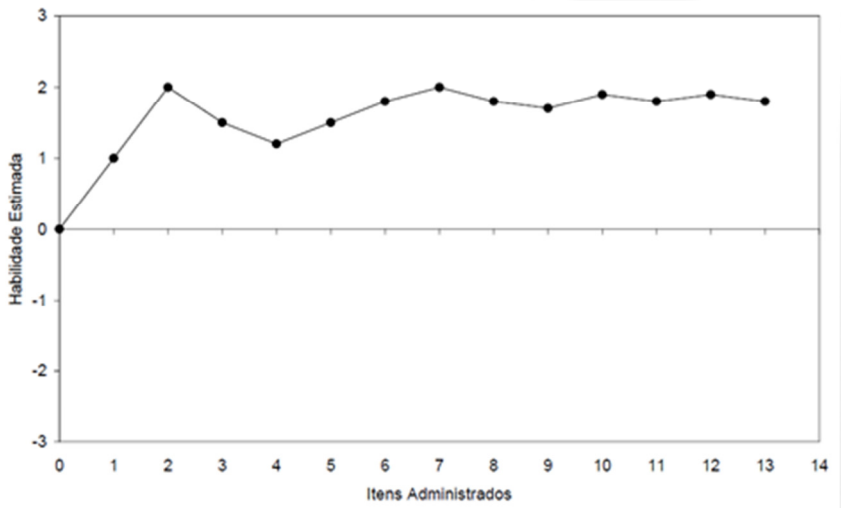
Fonte: Adaptado de Fetzer el tal., 2011

Uma ilustração de um exemplo típico de uso de um TAC é apresentada na Gráfico 2 onde nota-se que a cada questão respondida corretamente o nível de habilidade estimada do

ISBN 978-85-61115-09-8

indivíduo aumenta, ao passo que a cada questão respondida incorretamente o nível de habilidade estimada diminui e encontra um ponto de equilíbrio entre os níveis 1 e 2.

Gráfico2 – Exemplo de um TAC



Fonte: Moreira Junior, 2011

Do ponto de vista arquitetural um TAC necessita de um banco de itens; um critério para a escolha da primeira questão a ser apresentada para o examinando; um algoritmo de seleção de itens; um algoritmo de estimativa da habilidade; e um critério para determinar a finalização do teste.Entre todos os itens arquiteturais citados o BI é o mais sensível e o fator que representa, direta ou indiretamente, a maior desvantagem no uso de TAC, pois requer cuidado na elaboração de uma grande quantidade de questões e na escolha da metodologia de calibração inicial; apresenta potencial aumento dos custos por necessitar envolver profissionais com domínio de modelos estatísticos; necessita de um sistema computacional capaz de processar os cálculos de maneira rápida e segura; e envolve procedimentos criteriosos sempre que houver inclusão ou exclusão de novas questões no banco de itens podendo, conforme o caso, ser necessário refazer a calibração inicial dos itens (PITON-GONÇALVES, 2013).

1.3 GAI – Geração Automática de Itens

A geração automática de itens basicamente se resume a criação de várias questões com base em um modelo. No modelo são definidas variáveis cujos valores mudam a cada item que é gerado, obedecendo um conjunto de regras de definição e/ou restrição para o enunciado da questão, a alternativa correta e as alternativas incorretas. Se o modelo de item é criado a partir de uma questão já calibrada do BI, os itens gerados a partir do modelo já podem incorporar as informações psicométricas da questão original, eliminando assim a necessidade de nova calibragem do banco, situação em que os itens criados são chamados de itens isomorfos enquanto que seus modelos recebem o nome de modelos isomorfos (REVUELTA, 2000; BEJAR et al, 2003; MOREIRA JUNIOR, 2011; GIERL; LAI, 2013).

A possibilidade de se criar grandes quantidades de itens a partir de modelos estabelecidos, através do uso de computadores com softwares específicos para este fim, propicia o aumento de tamanho do BI e a diminuição da taxa de exposição dos itens em TACs, enquanto garante a equivalência do ponto de vista psicométrico quando da estimativa

ISBN 978-85-61115-09-8

do nível de proficiência de cada indivíduo. Contudo, estudos de simulação apresentados por Revuelta (2000) e Luecht (2013) demostraram que existe tendência de aumento no erro padrão da estimativa das proficiências e na estimativa dos parâmetros dos itens, em especial do parâmetro de discriminação, à medida que se aumenta a quantidade de itens isomorfos aplicados em um mesmo TAC, um efeito indesejado em testes de classificação e seleção, onde se exige alto grau de certeza na mensuração das habilidades. Todavia, o uso de itens isomorfos não deve causar impacto quando se trata de avaliações diagnósticas ou formativas já que tais avaliações não têm por finalidade classificar ou selecionar indivíduos, mas apenas fixar conteúdos (SOUZA, 2010).

Foulonneau & Ras (2013) descrevem um processo de engenharia para a geração automática de itens. Neste processo, apresentado na Figura 2, o passo Mapear Construto descreve as habilidades a serem medidas. O passo Modelar Evidência indica os conhecimentos necessários para diferentes níveis de performance, a serem refletidos em uma escala de pontuação de proficiência. Modelar Tarefas define a combinação de habilidades cognitivas e os objetos de conhecimento necessários para medir a proficiência em cada nível da escala. Modelar Itens estabelece o conjunto de habilidades e conhecimentos implícitos que são requeridos para a solução dos itens com base em tal modelo, além de conter as variáveis que devem ser resolvidas para a geração de itens. Resolver Variáveis, portanto resolve as variáveis e gera os itens que são armazenados em um BI, e por fim, a Entregar Item trata da aplicação do item em um teste.

Figura 2 – Processo de GAI

Fonte: Adaptado de Foulonneau&Ras, 2013

Uma lista de outros processos de GAI também pode ser obtida em Foulonneau & Ras (2013).

1.4 Dados abertos

Até há poucos anos a *Web* permitia a criação de um espaço virtual para compartilhamento de documentos que poderiam ser ou estar ligados a outros documentos não importando em qual lugar do globo estes ou aqueles estivessem hospedados. No atual contexto tecnológico chama-se esta realidade de Web de Documentos. O conceito *Linking Open Data* (LOD), visa transformar a Web de Documentos em uma Web de Dados (HAUSENBLAS; KARNSTEDT, 2010) provendo um novo paradigma baseado em tecnologias específicas para que tanto informação quanto dados brutos, contidos nos documentos, sejam acessíveis e interpretáveis por *softwares* de busca e processamento. Portanto, assim como a Web mudou a forma de como consumir documentos à procura de informações ou dados, o LOD procura mudar a forma de como descobrir, acessar, integrar e utilizar os dados ou informações.

O projeto LOD, que originou o conceito de mesmo nome, teve seu início dentro do Grupo de Educação e Extensão da Web Semântica do W3C, que a partir de março de 2008 repassou suas atividades para outros grupos independentes. A partir de então vários conjuntos de dados (*datasets*) foram publicados utilizando o formato de dados abertos sugerido pelo LOD, o *Linked Data*, que baseia-se na linguagem *Resource Description Framework* (RDF) (HEBELER, 2009).

ISBN 978-85-61115-09-8

A ideia essencial do modelo *Linked Data* através da linguagem RDF é demonstrar a ligação entre os dados através de triplas – um conjunto de Sujeito-Predicado-Objeto – utilizando o conceito de grafos e assim permitindo que determinado dado possa ser ligado a outro que fisicamente pode estar armazenado em outro arquivo, em qualquer outro lugar da *web*. Assim tem-se uma rede mundial de dados, passível de navegação. As ligações RDF, portanto, permitem navegar a partir de um item de dado dentro de uma fonte de dados para itens de dados relacionados dentro de outras fontes usando um navegador da Web Semântica. Ligações RDF também podem ser seguidas pelos *crawlers* dos motores de busca, que podem fornecer um resultado de busca mais sofisticado e outras formas de consulta sobre os dados rastreados a partir do resultado obtido. Como os resultados da consulta são dados estruturados e não apenas *links* para páginas HTML ou outros tipos de arquivos, eles podem ser usados dentro de outras aplicações de negócio de forma automatizada ou semiautomatizada.

Os três elementos de uma declaração ou tripla têm significados semelhantes aos seus significados nas gramáticas dos idiomas latinos. O sujeito de uma declaração é "coisa" que a declaração descreve, enquanto que o predicado descreve uma relação entre o sujeito e o objeto. Uma ilustração deste conceito pode ser verificada na Listagem 1.

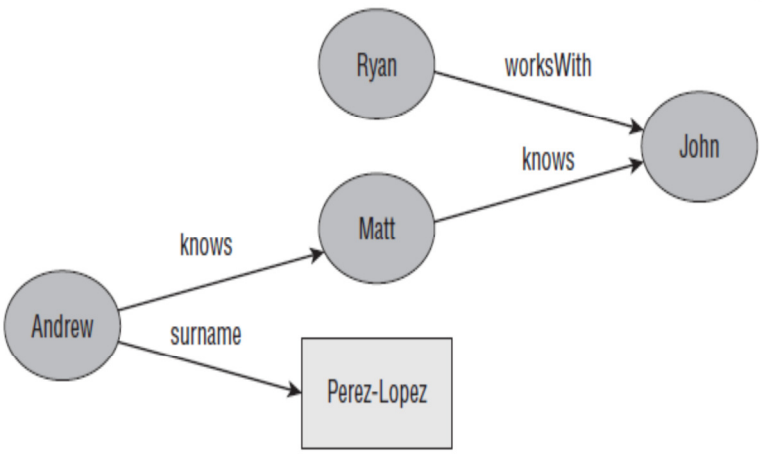
Listagem 1 – Informações sobre pessoas

```
Andrew knows Matt.
Andrew's surname is Perez-Lopez.
Matt knows John.
Ryan works with John.
```

Fonte: Hebeler et al., 2009

A Figura 3 é uma representação gráfica do conjunto de declarações apresentadas na Listagem 1. Afirmações nesta forma também são representadas em uma estrutura de grafo com sujeitos e objetos de cada declaração como nós, e predicados como bordas ou fronteiras. Este, portanto, é o modelo de dados LOD que é formalizado na linguagem RDF.

Figura 3 – Representação gráfica das sentenças apresentadas na Listagem 1



Fonte: Hebeler et al., 2009

Cada um dos elementos do grafo representam informações ou dados existentes em algum documento publicado na Internet, e assim cada um deles é representado na realidade

ISBN 978-85-61115-09-8

por uma *Uniform Resource Identifier* (URI) que são endereços desses elementos ou recursos na Internet e que fornecem a base para a infraestrutura de compartilhamento de dados no modelo LOD. Um exemplo do grafo apresentado pela Figura 3 serializado em formato RDF pode ser verificado na Listagem 2.

Listagem 2 – Conteúdo da Listagem 1 e Figura 3 serializado em RDF

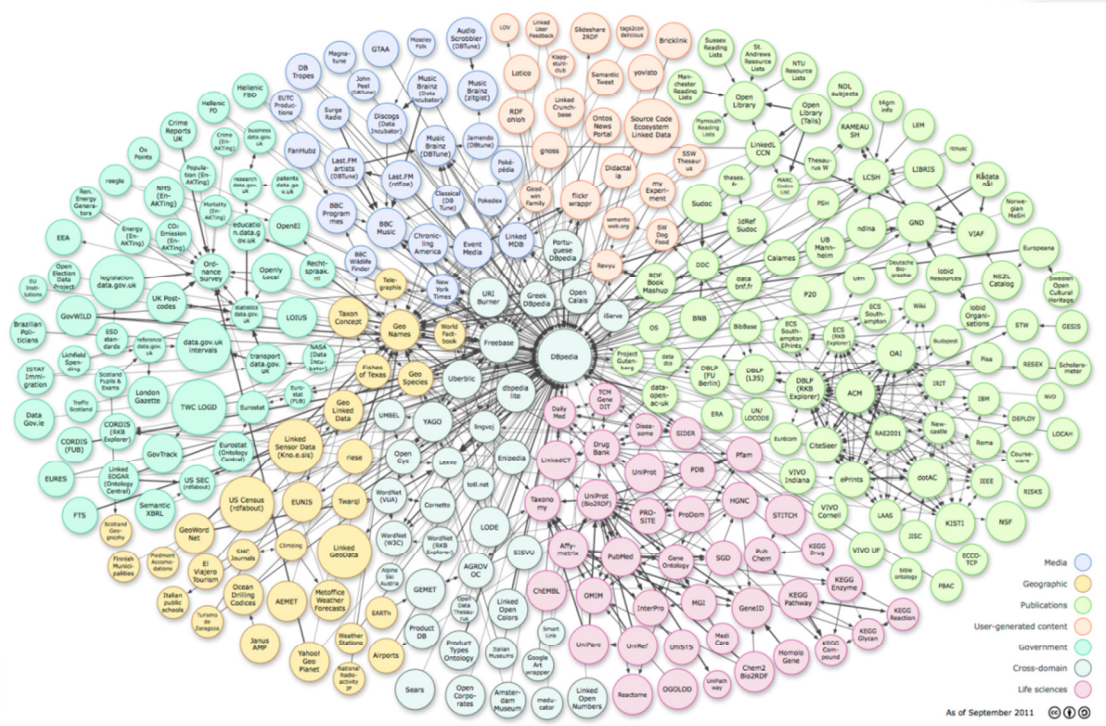
```
<rdf:RDF
    xmlns:people="http://semwebprogramming.net/people#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:ext="http://semwebprogramming.net/2008/06/ont/
      foaf-extension#">
<!-- This is a comment. -->
  <rdf:Description rdf:about="http://semwebprogramming.net/
       people#Ryan">
    <ext:worksWith
          rdf:resource="http://semwebprogramming.net/people#John"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://semwebprogramming.net/
       people#Matt">
    <foaf:knows
          rdf:resource="http://semwebprogramming.net/people#John"/>
  </rdf:Description>
  <rdf:Description
          rdf:about="http://semwebprogramming.net/people#Andrew">
    <foaf:surname>Perez-Lopez</foaf:surname>
    <foaf:knows
          rdf:resource="http://semwebprogramming.net/people#Matt"/>
  </rdf:Description>
</rdf:RDF>
```

Fonte: Hebeler et al., 2009

Neste exemplo, o arquivo RDF apresentado na Listagem 2 estaria disponível na *Internet* através da URL *<http://semwebprogramming.net/>* e acessível através de qualquer *software* que conheça sua existência e que consiga ler as informações existentes dentro dele. Desta maneira, tantos arquivos quanto se desejar criar, com descrição dos mais variados tipos de informações podem ser disponibilizados em qualquer servidor acessível pelos usuários da *Internet*. A Figura 4 mostra os conjuntos de dados que foram publicados e interligados pelo projeto LOD até setembro de 2011, data de sua última atualização. Coletivamente, os 295 conjuntos de dados consistem em mais de 31 bilhões de triplas RDF, que são interligadas por cerca de 504 milhões de ligações RDF (LINKING OPEN DATA, 2014).

ISBN 978-85-61115-09-8

Figura 4 – Diagrama da nuvem LOD



Fonte: LOD Cloud, 2014

Atualmente, a plataforma DataHub, responsável por manter o catálogo dos *datasets* disponíveis na *Internet*, apresenta uma listagem de 9.855 *datasets* sobre os mais variados assuntos (DATAHUB, 2014). Cada nó apresentado na Figura 4, e cada um dos mais de nove mil *datasets* listados pelo DataHub, uma vez disponibilizados, permitem o acesso, consulta e recuperação dos dados, normalmente através de pelo menos uma de três formas possíveis: um *software* do tipo *browser*, chamado *browser* semântico; um *site* de buscas semântico; e através de consultas estruturadas através da linguagem SPARQL, muito semelhante à linguagem SQL utilizada para manipulação de dados em bancos de dados relacionais.

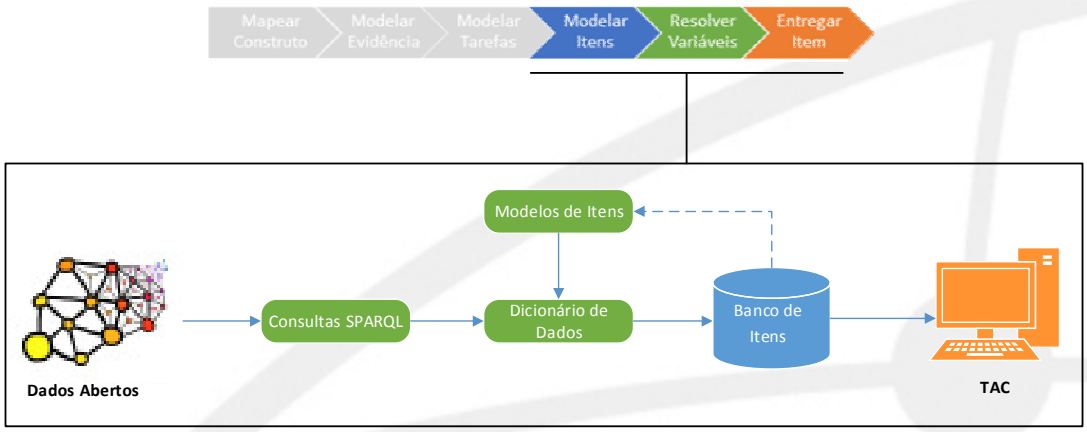
A proposta de GAI deste trabalho baseia-se no acesso ao *dataset* DBpedia com o uso de consultas SPARQL.

2 Estudo de caso

A partir do processo de GAI apresentado na seção 1.3, a pesquisa em que este trabalho está inserido concentra-se nos passos Modelar Itens, Resolver Variáveis e Entregar Item com acesso a base de LOD DBpedia para geração de itens isomorfos, conforme apresentado na Figura 5. Mais especificamente em relação ao presente estudo, concentra-se nopasso de Resolver Variáveis.

ISBN 978-85-61115-09-8

Figura 5 – Escopo do sistema de GAI



Para a realização deste estudo são propostos dois modelos de itens encontrados facilmente nas seções de exercícios de fixação em livros didáticos de Geografia e Língua Portuguesa referentes a conteúdos do Ensino Fundamental. São eles:

a) Qual a capital do estado de <*nome do estado*>?
b) Entre os poetas listados abaixo qual pertence ao movimento artístico <*nome do movimento*>?

Para o uso de dados abertos no preenchimento das variáveis definidas no modelo de item é necessário o domínio de alguns padrões e tecnologias como XML, IMS-QTI, RDF, OWL e SPARQL.

O padrão IMS-QTI fornece uma especificação para possibilitar a troca de itens, testes e dados de resultados entre ferramentas de autoria, bancos de itens, ferramentas de construção de testes, ambientes de aprendizagem e sistemas para avaliações (IMS GLOBAL, 2014). Este padrão é serializado em formato XML e possui uma série de identificadores únicos para representar o enunciado da questão, cada uma de suas alternativas, elementos multimídia que devem ser agregados, parâmetros psicométricos, entre outros. Na Listagem 3 e na Listagem 4 é possível visualizar trechos de um arquivo XML utilizando o padrão IMS-QTI onde se descreve os modelos de itens formulados para este estudo.

Listagem 3 – Conteúdo de arquivo XML com a descrição do primeiro modelo de item

```
<choiceInteraction responseIdentifier="RESPONSE" shuffle="false" maxChoices="1">
<prompt>Qual a capital do estado de/do {estado}?</prompt>
<simpleChoice identifier="{codigoResposta1}">{textoResposta1}</simpleChoice>
<simpleChoice identifier="{codigoResposta2}">{textoResposta2}</simpleChoice>
<simpleChoice identifier="{codigoResposta3}">{textoResposta3}</simpleChoice>
<simpleChoice identifier="{codigoResposta4}">{textoResposta4}</simpleChoice>
</choiceInteraction>
```

Os identificadores são representados entre os sinais <> e as variáveis a serem resolvidas entre os sinais *{}*. O identificador *choiceInteraction* representa uma questão que deve ser apresentada a algum examinando, através do identificador *prompt*, com um conjunto de opções, cada qual definida por um identificador *simpleChoice*. A tarefa do examinando é o de selecionar uma ou mais das opções disponíveis, até um máximo definido por*maxChoices*.

ISBN 978-85-61115-09-8

Listagem 4 – Conteúdo de arquivo XML com a descrição do segundo modelo de item

```
<choiceInteraction responseIdentifier="RESPONSE" shuffle="false" maxChoices="1">
<prompt>Entre os poetas listados abaixo qual pertence ao movimento artístico
{nomeMovimento}?</prompt>
<simpleChoice identifier="{codigoResposta1}">{textoResposta1}</simpleChoice>
<simpleChoice identifier="{codigoResposta2}">{textoResposta2}</simpleChoice>
<simpleChoice identifier="{codigoResposta3}">{textoResposta3}</simpleChoice>
<simpleChoice identifier="{codigoResposta4}">{textoResposta4}</simpleChoice>
</choiceInteraction>
```

Como um passo preliminar para Resolver Variáveis é necessário obter os dados de um *dataset*, o DBpedia neste caso, através da realização de consultas com a linguagem SPARQL sobre dados serializados no formato RDF normalmente organizados através de alguma ontologia. O comando de consulta pode ser testado através de um *SPARQL endpoint*, uma espécie de mecanismo de busca cuja interface permite a escrita de comandos dessa linguagem, que é disponibilizado pelo próprio *dataset*, neste caso através do endereço <*http://dbpedia.org/sparql*>. Exemplo de uma consulta SPARQL que busca por uma lista de estados brasileiros juntamente com suas capitais que é utilizado para gerar os itens para o modelo apresentado na Listagem 3, é apresentado na Listagem 5, enquanto que o modelo da Listagem 4 tem sua respectiva consulta apresentada na Listagem 6.

Listagem 5– Consulta SPARQL para recuperar a lista de estados brasileiros e suas capitais

```
PREFIX dbpedia-type: <http://dbpedia.org/class/yago/>
PREFIX dbpedia-prop: <http://dbpedia.org/property/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?estado ?capital
WHERE  {
    ?itemEstado a dbpedia-type:StatesOfBrazil ;
                rdfs:label ?estado ;
                dbpedia-prop:seat ?itemCapital.
    ?itemCapital rdfs:label ?capital .
    FILTER (
        langMatches( lang(?estado), "PT" ) &&
        langMatches( lang(?capital), "PT" )
)
}
```

O resultado retornado pela consulta da Listagem 5 foi um conjunto com 24 estados: Acre, Alagoas, Amapá, Amazonas, Bahia, Ceará, Distrito Federal, Espírito Santo, Goiás, Maranhão, Mato Grosso, Mato Grosso do Sul, Pará, Paraíba, Paraná, Pernambuco, Piauí, Rio Grande do Norte, Rio Grande do Sul, Rondônia, Roraima, São Paulo, Sergipe e Tocantins. Não foram apresentados no resultado os estados de Minas Gerais e Rio de Janeiro devido ao predicado <*http://dbpedia.org/property/seat*> possuir um valor literal e não o endereço de outro sujeito no *dataset*, e Santa Catarina devido a inexistência da ligação do predicado <*http://dbpedia.org/property/seat*> definido para si. O resultado obtido ao se consultar especificamente pelo estado do Rio de Janeiro retornou resultado do predicado <*http://dbpedia.org/property/seat*>o valor “20”, o que desqualificaria este item caso ele estivesse presente no resultado original da consulta.

ISBN 978-85-61115-09-8

Listagem 6– Consulta SPARQL para recuperar a lista de poetas brasileiros e os movimentos artísticos dos quais participaram

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX cat: <http://dbpedia.org/resource/Category:>
SELECT ?nomePoeta ?nomeMovimento
WHERE{
    ?poeta dcterms:subject cat:Brazilian_poets;
rdfs:label ?nomePoeta ;
           dbo:movement ?movimento .
    ?movimento rdfs:label ?nomeMovimento
    FILTER (
        LANG(?nomePoeta) = "pt" &&
        LANG(?nomeMovimento) = "pt"
    )
}
```

O resultado retornado pela consulta da Listagem 6 foi um conjunto de 50 poetas brasileiros com seus respectivos movimentos artísticos dos quais foram participantes. Neste conjunto de resultados foram identificados 4 resultados inválidos nos quais não se identificava corretamente o nome do movimento artístico: “Poesia concreta”, “Poesia fonética”, “Novas mídias” e “Literatura do Brasil”. Outros resultados foram corretamente identificados em cada movimento artístico: Barroco com 1 resultado, Parnasianismo com 11, Modernismo com 3, Romantismo com 20, Neoclassicismo com 7, Naturalismo com 1, Neorromantismo com 1 e Simbolismo com 2 resultados. O resulto de apenas 1 poeta barroco foi investigado através da comparação da mesma pesquisa com o uso de outras ontologias, que trouxeram resultados corretos, mas com poetas diferentes, por exemplo ao se trocar o objeto *cat:Brazilian_poets* por *<http://dbpedia.org/class/yago/BrazilianPoets>* o conjunto de itens retornados para Barroco continuou sendo 1, mas ao invés de Gregório de Matos da consulta original obteve-se Bento Teixeira com a nova consulta.

Uma consulta SPARQL recupera os dados relacionados ao recurso pesquisado normalmente na forma de uma resposta correta. Contudo, na formulação de itens de múltipla escolha, como apresentado na Listagem 3 e na Listagem 4, são necessários também algumas alterativas incorretas, a que se dá o nome de distratores. Sendo assim, pode-se escolher aleatoriamente outros nomes de capitais ou poetas respectivamente para as outras alternativas incorretas de acordo com o modelo de item. Por fim, a partir do modelo de item e dos dados retornados pelas consultas SPARQL o sistema de GAI gera um conjunto de questões que são armazenadas em um BI podendo incluir informações extras indicando a base utilizada para extração dos dados, data da extração e geração, entre outras.Todos os resultados obtidos pela consulta SPARQL podem também ser armazenados localmente utilizando o formato XML ou em um banco de dados relacional.

A partir da leitura automatizada dos resultados retornados pelas consultasfoi possível gerar os itens preenchendo as variáveis de seus respectivos modelos através de análise combinatória. Para o modelo de item apresentado na Listagem 3 foi possível a construção de 1771 itens de teste diferentes em função das possibilidades de arranjo do conjunto de itens distratores sendo que para cada pergunta sobre um dos estados recuperados foram encontradas 77 possibilidades de arranjo. Para o modelo sugerido na Listagem 4 foi possível a construção de 49423 itens diferentes em função da possibilidade de arranjos de subconjuntos de poetas sem aqueles pertencentes a cada movimento artístico cuja questão era gerada e também considerando que alguns poetas participaram de mais de um movimento artístico e, portanto, seus nomes deveriam ser excluídos antes da escolha aleatória dos itens distratores.

ISBN 978-85-61115-09-8

Considerações Finais

Mesmo considerando que este trabalho apresenta o relato inicial de nossa pesquisa sobre Geração Automática de Itens para avaliações educacionais de baixo risco com o uso de dados abertos algumas considerações já podem feitas.

Percebe-se a possibilidade prática da GAI utilizando as tecnologias descritas já que a partir do modelo de item da Listagem 3 foi possível construir 1771 itens, enquanto para o modelo da Listagem 4 foi possível construir 49423 itens.

Também é possível identificar duas situações sensíveis através dos resultados obtidos pelas consultas SPARQL: a primeira quando os dados obtidos podem ser incorretos como os 4 itens identificados pela consulta da Listagem 4, ou a segunda quando os dados podem estar ausentes ou não ligados através da ontologia utilizada na consulta como é o caso da apresentação de apenas um poeta do movimento Barroco, Gregório de Matos, e ao se trocar a ontologia utilizada também se recebe apenas um poeta, mas Bento Teixeira.

É necessária a identificação prévia dos *datasets* e seus domínios de conhecimento cujos dados estão publicados no formato LOD, bem como das diversas ontologias e vocabulários existentes, para que se possa realizar uma adequada extração de dados através de uma ferramenta de autoria por exemplo. Sugere-se o acesso ao *sitehttp://datahub.io/*.

Os modelos de itens utilizados inicialmente, como o apresentado na Listagem 3 e na Listagem 4, foram bastante simples. Tem-se em vista, para trabalhos futuros, a identificação de padrões para a criação de itens mais complexos, como aqueles utilizados no Exame Nacional do Ensino Médio (ENEM) e outros exames de larga escala semelhantes.

Outro avanço esperado para esta pesquisa é a criação de uma ferramenta de autoria com foco, em um primeiro momento, na geração semiautomática de itens para avaliações de baixo risco, onde usuários especialistas em conteúdo, mas sem conhecimento técnico das tecnologias envolvidas, possam criar modelos de itens e dar início ao processo de geração dos itens, com posterior revisão e exclusão de itens inadequados.

Por fim, para a validação final da qualidade dos itens gerados, espera-se realizar a entrega do item através de TAC baseado na TRI com o modelo ML3, pesquisa que tem sido desenvolvida em paralelo a este trabalho e que encontra-se em estágio mais adiantado.

Referências

ALVES, D. T. et al. **Análise de metodologia baseada no sistema de ensino individualizado de Keller aplicada em um curso introdutório de eletromagnetismo**. Revista Brasileira de Ensino de Física, São Paulo, v. 33, n. 1, Mar. 2011. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-11172011000100014&lng=en&nrm=iso>. Acesso em 07 jul. 2013.

BEJAR, Isaac I. et al. **A feasibility study of on-the-fly item generation in adaptive testing**. The Journal of technology, learning and assessment, v. 2, n. 3, 2003.

BIRNBAUM, A. Some Latent Trait Models and Their Use in Infering an Examinee's Ability. In: LORD, F. M.; NOVICK, M. R. **Statistical Theories of Mental Test Scores**. Reading, MA: Addison-Wesley, 1968.

COSTA, D. R. **Métodos Estatísticos em Testes Adaptativos Informatizados**. Dissertação. 2009. 120 f. Dissertação (Mestrado) – Departamento de Métodos Estatísticos, Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

DATAHUB. **DataHub**. Disponível em: <http://datahub.io/>. Acesso em: 13 set. 2014

ISBN 978-85-61115-09-8

FETZER, M. et al. **Computer Adaptive Testing (CAT) in an Employment Context**. White paper. Roswell, USA: PreVisor, 2011.

FOULONNEAU, M.; RAS, E. Assessment Item Generation, the way forward. In: **2013 International Computer Assisted Assessment (CAA) Conference**, 20., 2013, Southampton, UK. Proceedings… Southampton, 2013..

GARCIA-CASTRO, R.; FENSEL, D.; ANTONIOU, G. (Eds.). **The Semantic Web**: ESWC 2011 Workshops. Springer, 2012.

GIERL, M. J.; LAI, H. Using Weak and Strong Theory to Create Item Models for Automatic Item Generation. In: GIERL, M. J.; HALADYNA, T. M. (Eds.). **Automatic item generation: Theory and practice**. Routledge, 2012.

HAUSENBLAS, M.; KARNSTEDT, M. Understanding Linked Open Data as a Web-Scale Database. In: **International Conference Nn Advances In Databases Knowledge And Data Applications (DBKDA)**, 2., 2010, Menuires. Proceedings... Menuires: IEEE, 2010.

HEBELER, J. et al. **Semantic Web Programming**. Indianapolis: Wiley Publishing, 2009.

IMS GLOBAL. **IMS Global Learning Consortium**. Disponível em: <http://www.imsglobal.org/>. Acesso em: 07 set. 2014.

JOHNSON, L. et al. **NMC Horizon Report:** 2013 Higher Education Edition. Austin, Texas: The New Media Consortium, 2013.

LINKING OPEN DATA. **The Linking Open Data Cloud**. Disponível em: <http://lod-cloud.net/>. Acessado em: 13 set. 2014.

LORD, F. M. **A theory of test scores** (No. 7). Psychometric Monograph, 1952.

LUECHT, R. M. An Introduction to Assessment Engineering for Automatic Item Generation. In: GIERL, M.; HALADYNA, T. M. (Eds). **Automatic Item Generation**: theory and practice. Taylor & Francis, 2013.

MOREIRA JUNIOR, F. J. **Sistemática para a implantação de Testes Adaptativos Informatizados baseados na Teoria da Resposta ao Item**. 2011, 334 f. Tese (Doutorado) – Centro Tecnológico, UFSC, Florianópolis.

PASQUALI, L. **Psicometria**: Teoria dos testes na Psicologia e na Educação. 4ª ed. Petrópolis: Vozes, 2011.

PITON-GONÇALVES, J. **Desafios e perspectivas da implementação computacional de Testes Adaptativos Multidimensionais para avaliações educacionais**. 2012, 153 f. Tese (Doutorado) – Instituto de Ciências Matemáticas e de Computação, ICMC/USP, São Carlos.

RASCH, G. **Probabilistic Models for Some Intelligence and Attainment Tests**. Copenhagen: Danish Institute for Educational Research, 1960.

REVUELTA, J. Estimación de habilidad mediante ítems isomorfos. Efectos en la fiabilidad de las puntuaciones. **Psicothema**, v. 12, n. 2, p. 303-307, 2000.

SCHEUERMANN, F,; BJÖRNSSON, J. (Eds.). **The Transition to Computer-Based Assessment:** New Approaches to Skills Assessment and Implications for Large-scale Testing. Luxemburgo: European Communities, 2009.

SOUZA, S. Z. Avaliação: da pedagogia da repetência à pedagogia da concorrência? In: DALBEN, A. I. L. F. et al. (Org). **Didática: convergências e tensões no campo da formação e do trabalho docente**. Belo Horizonte: Autêntica, 2010.

ISBN 978-85-61115-09-8

THOMPSON, N. A.; WEISS, D. J. A Framework for the Development of Computerized Adaptive Tests. **Practical Assessment, Research & Evaluation**, 16(1), 2011. Disponível em: <http://pareonline.net/getvn.asp? v=16&n=1>. Acesso em 10 jul. 2013.

VAN DER LINDEN, W. J.; GLAS, C. A. W. **Elements of Adaptive Testing**. Statistical for Social and Behavioral Sciences. New York: Springer Science+Business Media, LLC, 2010.

WRIGHT, B. D. **Sample-free test calibration and person measurement**. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: ETS - Educational Testing Service, 1968.

ISBN 978-85-61115-09-8

UM EXPERIMENTO ENVOLVENDO A GERAÇÃO DE MAPAS DE TÓPICOS AUTOMATIZADA A PARTIR DOS DADOS ABERTOS DO SISTEMA DE CONVÊNIOS (SICONV)

Mateus Lohn Andriani
mtslohn@gmail.com

Flavio Ceci
flavio.ceci@unisul.br

Denilson Sell
denilson@stela.org.br

José Leomar Todesco
tite@egc.ufsc.br

Resumo
Este artigo apresenta um experimento envolvendo a geração de mapas de tópicos a partir de dados abertos governamentais, aplicada nesse cenário aos dados do Sistema de Convênios (SICONV) do governo federal brasileiro. A abordagem empregada no experimento utiliza as técnicas de reconhecimento de entidades e de geração de matrizes de correlação para a descoberta de conhecimento. O experimento foi conduzido na forma de um estudo de caso dentro da Plataforma Aquarius, a qual concentra informações analíticas sobre diversos dados governamentais. Os mapas de tópicos gerados permitiram a visualização sumarizada das principais palavras-chaves mencionadas e sua interligação com outras palavras-chaves relacionadas aos dados descritivos dos convênios, sendo que a abordagem pode ser replicada para outros conjuntos de documentos.

Palavras-chave: Dados Abertos Governamentais.Descoberta de Conhecimento. Mineração de Texto.

Abstract
This article presents an experiment involving the generation of topic maps from open government data, applied in this scenario to the data of the Covenants System (SICONV) of the Brazilian federal government. The approach used in the experiment uses the entities recognition and correlation matrix generation techniques for knowledge discovery. The experiment was conducted as a case study within the Platform Aquarius, which concentrates analytical information of various government databases. The generated topic maps allowed the summarized view of the main keywords mentioned and its interconnection with other keywords related to the descriptive data of the covenants, wherein the approach can be replicated to other sets of documents.

Key Words: Open Government Data.Knowledge Discovery. Text Mining.

ISBN 978-85-61115-09-8

Introdução

A informação é um recurso fundamental para o processo de tomada de decisão, permitindo o levantamento de problemas dentro de quaisquer cenários e suas possíveis soluções. Fatores como a quantidade e a qualidade de informações influenciam diretamente nos resultados obtidos através das análises. Com o avanço da tecnologia computacional, cada vez mais volumes maiores de informação são captados e processados, com o objetivo de detectar esses problemas e indicar pistas para suas resoluções.

Uma grande quantidade das informações registradas não possui conteúdo computacionalmente estruturado (tais como esse texto, que para um computador pode não passar de uma série de caracteres contidos dentro de um arquivo). Muitos dos sistemas não conseguem discernir ou validar a informação ali contida; no entanto, para o ser humano, a leitura da informação não estruturada produz significado através da interpretação do conteúdo. De acordo com Grimes (2008), cerca de 80% a 85% das informações de uma organização estão contidas em documentos de texto. Dessa forma, ao se trabalhar somente com dados estruturados, o universo de análise é reduzido para até 15% do total de informações de uma organização. Esforços para extrair mais valor dessa informação não interpretável de forma simplificada pelos computadores constituem grande parte das pesquisas na área de mineração de texto (também nomeada *Knowledge Discovery in Text*, ou KDT).

A maior parte das iniciativas ligadas ao movimento *Open Government* está ligada a apresentação de informações quantitativas, derivadas de técnicas de mineração de dados (também nomeada *Knowledge Discovery in Databases*, ou KDD). Apesar da acurácia que esse tipo de técnica possui, o universo de análise torna-se limitado, visto que parte das informações inseridas em sistemas governamentais são armazenadas no formato de texto não estruturado. O uso de técnicas de KDT poderia permitir, por exemplo, identificar o modelo de abordagem do governo, através da mineração das motivações associadas aos investimentos realizados e das áreas nas quais os investimentos estão sendo realizados.

Esse artigo apresenta uma abordagem para a produção de mapas de tópicos a partir de um conjunto de informações não estruturadas, aplicando-a como experimento a uma subparte dos dados abertos disponibilizados pelo governo federal do Brasil, utilizando-se de técnicas de mineração de texto. A subparte adotada foi a dos dados do Sistema de Convênios (SICONV), disponibilizado pela Infraestrutura Nacional de Dados Abertos (INDA). O resultado final do experimento é a geração de um mapa de tópicos, que ilustra as palavras-chaves interligadas de acordo com sua coocorrência dentro dos projetos de convênio realizados pelo governo.

O experimento foi desenvolvido dentro da Plataforma Aquarius, que possui como objetivo principal transparecer os dados governamentais do Ministério da Ciência, Tecnologia e Inovação (MCTI).

O artigo segue com a conceitualização das técnicas de mineração de texto aplicadas na abordagem (reconhecimento de entidades e geração de matrizes de correlação) e o embasamento da definição de convênio dentro da legislação brasileira, que vem a ser a fonte de informação utilizada para a aplicação da abordagem. Na sequência, é apresentada a abordagem e o estudo de caso no qual foi conduzido o experimento. Por fim, as considerações finais sobre o processo são descritas.

ISBN 978-85-61115-09-8

1 Mineração de texto e descoberta de conhecimento

Segundo Tan (2012), a descoberta de conhecimento pode ser definida como a área da Ciência da Computação que trabalha com a pesquisa automatizada sobre grandes conjuntos de dados visando encontrar padrões que podem ser considerados ativos de conhecimento. A área utiliza diversas técnicas e procedimentos computacionais, sendo um subconjunto dessas os processos de mineração e processamento de texto.

A mineração de texto refere-se ao processo de extração de conhecimento a partir de documentos de texto. É uma área derivada da mineração de dados tradicional, que é executada sobre bases de dados estruturadas.

A área desperta grande interesse, pois a mesma pode proporcionar a descoberta de mais ativos de conhecimento em relação à mineração de dados tradicional. No entanto, este tipo de mineração possui um alto grau de complexidade, visto que as informações existentes nestes documentos estão muitas vezes desestruturadas e imprecisas, em razão de sua própria natureza.

As subseções seguintes tratam das técnicas de mineração de texto empregadas na abordagem para a geração de mapas de tópicos. No contexto desse trabalho, são utilizados os procedimentos de reconhecimento de entidades e de geração de matrizes de correlação.

1Reconhecimento de entidades

O reconhecimento de entidades consiste na detecção de elementos informativos em textos, tais como nomes de pessoas, nomes de empresas, cidades, datas e valores monetários. Um exemplo do trabalho realizado no reconhecimento de entidades pode ser visto no Quadro 1.

Quadro 1 - Exemplo de aplicação da técnica de reconhecimento de entidades

Texto original	O oficial da ONU Erkeus dirige-se para Bagdá.
Texto com os marcadores	O oficial da [ONU] [Erkeus] dirige-se para [Bagdá].
Entidades extraídas (com a classe entre parênteses)	ONU (organização), Erkeus (pessoa) e Bagdá (localização geográfica)

Fonte: Adaptado de Sang e De Meulder (2003)

Esta técnica é de grande auxílio na extração e recuperação de informação. O reconhecimento de entidades pode ser aplicado no pré-processamento do texto, fornecendo entidades relevantes para serem analisadas nos documentos. (MIKHEEV; MOENS; GROVER, 1999; KOZAREVA; FERRÁNDEZ; MONTOYO, 2012)

Os primeiros sistemas de reconhecimento de entidades desenvolvidos utilizavam como base ferramentas linguísticas e índices geográficos, o que tornava o desenvolvimento difícil e acoplado ao domínio de extração. (KOZAREVA; FERRÁNDEZ; MONTOYO, 2012). No entanto, a importância do tamanho do índice utilizado foi contestada por Krupka e Hausman (1998). No trabalho desenvolvido por eles, uma série de índice foi utilizada para testes de eficiência, onde enquanto um índice com 110.000 verbetes auxiliava a ferramenta a fornecer o aproveitamento de 91,6% no reconhecimento de entidades, a utilização de um índice com apenas 9.000 verbetes traz o aproveitamento de 89,1%.

ISBN 978-85-61115-09-8

Os sistemas de reconhecimento de entidades mais recentes utilizam técnicas de aprendizado de máquina para recuperar os termos a partir dos textos. Nadeau e Sekine (2007) descrevem os tipos de algoritmos de reconhecimento de entidades de acordo quanto ao formato de aprendizado:

a) Aprendizado supervisionado: neste formato, o algoritmo considera como entidades somente os termos marcados em um *corpus*[14].
b) Aprendizado semi-supervisionado: neste formato, o algoritmo considera como entidades os termos marcados em um *corpus* e os termos que se encaixam no mesmo padrão de utilização das entidades do *corpus* em cada uma das frases dos documentos analisados. O *corpus* de entidades funciona como um ponto de partida para o reconhecimento de entidades.
c) Aprendizado não-supervisionado: neste formato, o algoritmo considera como entidades os termos dos documentos que seguem algum padrão de construção que seja considerado válido pelo mesmo. Entre os padrões de validação estão a similaridade de contexto ou a construção de acordo com um padrão léxico. O algoritmo ainda pode validar as entidades através de técnicas de estatística sobre um número grande de documentos.

1.2 Geração de matrizes de correlação

Segundo Tabachnick e Fidell (2007), uma das formas mais comuns de representar o relacionamento entre variáveis é através de uma matriz de correlação.

A matriz de correlação caracteriza-se por ser uma matriz quadrada, onde cada linha e cada coluna representam uma variável de análise. Através da intersecção de uma linha e uma coluna, é possível obter os dados relativos à coocorrência das variáveis. Alguns dos dados que podem ser armazenados em cada célula desta matriz são a frequência conjunta e o coeficiente de correlação das variáveis. (TABACHNICK; FIDELL, 2007; PECK; DEVORE, 2012)

Uma matriz de correlação é derivada de um modelo onde as variáveis são associadas a cada elemento agrupador. Este modelo pode ser representado através de uma matriz, na qual cada documento apresenta o número de frequências em cada elemento para cada uma das variáveis selecionadas para análise. O Quadro 2 apresenta a matriz resultante de uma análise da frequência de três variáveis dentro de um conjunto de nove elementos.

Quadro 2 – Matriz variável-elemento

Elemento	X1	X2	X3
E1	1	0	1
E2	1	1	0
E3	0	1	0
E4	0	0	1
E5	0	0	1
E6	1	0	0

[14]Coletânea de documentos ou textos sobre de um determinado assunto de interesse, construído para análise de fenômenos (Dicionário Priberam da Língua Portuguesa [em linha], 2014).

ISBN 978-85-61115-09-8

E7	1	1	0
E8	0	0	1
E9	0	0	1

Fonte: Elaboração dos autores (2014)

A partir da matriz variável-elemento, é possível verificar quais variáveis ocorrem em um mesmo elemento. Esta informação é a existência de frequência conjunta, que pode ser obtida para todas as combinações possíveis de variáveis de análise. Neste trabalho, será utilizada a frequência conjunta de todas as combinações de duas variáveis para cada elemento.

A partir da sumarização da frequência conjunta para cada uma das variáveis, é possível montar a matriz de correlação. O Quadro 3 apresenta a matriz de correlação gerada a partir da matriz variável-elemento.

Quadro 3 – Matriz de correlação entre variáveis (apresentando a frequência conjunta)

Elemento	**X1**	**X2**	**X3**
X1	4	2	1
X2	2	3	0
X3	1	0	5

Fonte: Elaboração dos autores (2014)

Na matriz de correlação, cada linha e cada coluna representam uma das variáveis de análise. A intersecção de uma linha com uma coluna apresenta o número de frequências conjuntas das variáveis selecionadas.

As matrizes de correlação são simétricas sobre a diagonal principal, visto que o número de frequência conjunta das variáveis não é alterado pela ordem de combinação. Desta forma, os resultados da diagonal inferior esquerda são refletidos também na diagonal superior direita (de acordo com os pares correspondentes). A combinação da variável pela própria variável traz a frequência individual da mesma, já que se trata da coocorrência da mesma variável. Desta forma, é uma pratica comum apresentar apenas a metade correspondente a diagonal superior direita ou a diagonal inferior esquerda, eliminando assim a redundância existente na matriz. (TABACHNICK; FIDELL, 2007)

As técnicas de geração de matrizes de correlação e de reconhecimento de entidades nomeadas foram aplicadas sobre os dados de convênios disponibilizados pelo governo federal. Com o intuito de contextualizar o objeto de análise, a próxima seção define o conceito de convênio e descreve como os dados podem ser recuperados, de forma que o processo apresentado neste artigo possa ser replicado para outras aplicações.

2 Convênios

Os convênios são instrumentos formais que disciplinam a transferência de recursos da União para entidades públicas (tais como Estados, Municípios ou o Distrito Federal) ou particulares, visando à execução de atividades de interesse recíproco e com cooperação mútua. (BRASIL, 2010)

O Decreto nº 6.170 apresenta a definição de convênio dentro da legislação brasileira:

ISBN 978-85-61115-09-8

> "Acordo, ajuste ou qualquer outro instrumento que discipline a transferência de recursos financeiros de dotações consignadas nos Orçamentos Fiscal e da Seguridade Social da União e tenha como partícipe, de um lado, órgão ou entidade da administração pública federal, direta ou indireta, e, de outro lado, órgão ou entidade da administração pública estadual, distrital ou municipal, direta ou indireta, ou ainda, entidades privadas sem fins lucrativos, visando a execução de programa de governo, envolvendo a realização de projeto, atividade, serviço, aquisição de bens ou evento de interesse recíproco, em regime de mútua cooperação" (DECRETO N. 6.170, 2007)

Os convênios diferenciam-se dos contratos convencionais. Os primeiros se caracterizam pelo interesse de ambas as partes na execução das atividades ou projetos em pauta e na cooperação mútua para o alcance das metas e objetivos propostos. Em contrapartida, nos contratos, o interesse das partes envolvidas é diverso, onde interessa unicamente à parte contratante a realização das atividades ou projetos e, à parte contratada, a remuneração correspondente à execução destas tarefas, não existindo a necessidade de cooperação entre as partes. (BRASIL, 2010)

Os repasses de recursos realizados através dos convênios são denominados transferências voluntárias. Estas transferências são realizadas a título de cooperação, auxílio ou assistência financeira e abrangem concedentes e convenentes, que são definidos conforme a Lei nº 10.707 de 30 de julho de 2003, que se refere às Diretrizes Orçamentárias (BRASIL, 2003; BRASIL, 2009):

a) Concedente: órgão ou entidade da administração pública direta ou indireta responsável pela transferência de recursos financeiros ou descentralização de créditos orçamentários destinados à transferência voluntária;
b) Convenente: órgão ou entidade da administração pública direta ou indireta dos governos estaduais, municipais ou do Distrito Federal, com o qual a administração federal pactue a execução de programa, projeto, atividade ou evento de duração certa, com recursos provenientes de transferência voluntária.

Os convênios podem ser originados a partir de emendas apresentadas ao Orçamento Fiscal da União (por deputado federal ou senador), propostas ou projetos formulados pelo próprio interessado (diretamente ao ministério ou à entidade que disponha de recursos aplicáveis ao objeto pretendido) ou de pedidos do próprio ministério ou entidade (ao detectarem a existência de necessidades ou carências ou ao implantar programas de governo). (BRASIL, 2003)

O Sistema de Gestão de Convênios e Contrato de Repasses (SICONV) centraliza os dados dos convênios e contratos de repasse realizados pelo Governo Federal. O sistema entrou em vigor em 1º de setembro de 2008 e é uso obrigatório para a realização das transferências voluntárias. De acordo com a Portaria nº 127/2008, todos os convênios firmados a partir do dia 29 de maio de 2008 e em estado de vigência no dia 31 de dezembro de 2009 possuem cadastro no sistema. (MINISTÉRIO DO PLANEJAMENTO, ORÇAMENTO E GESTÃO, 2009)

Os dados dos convênios estão disponibilizados para *download* através do portal de dados abertos do SICONV, sendo atualizados diariamente com as novas informações que são cadastradas no sistema. Além dos dados, é disponibilizada também uma API (*Application Layer Interface*) utilizando o padrão REST, que disponibiliza os dados de um convênio em específico ou um agrupamento de convênios nos formatos HTML, JSON, CSV e XML.

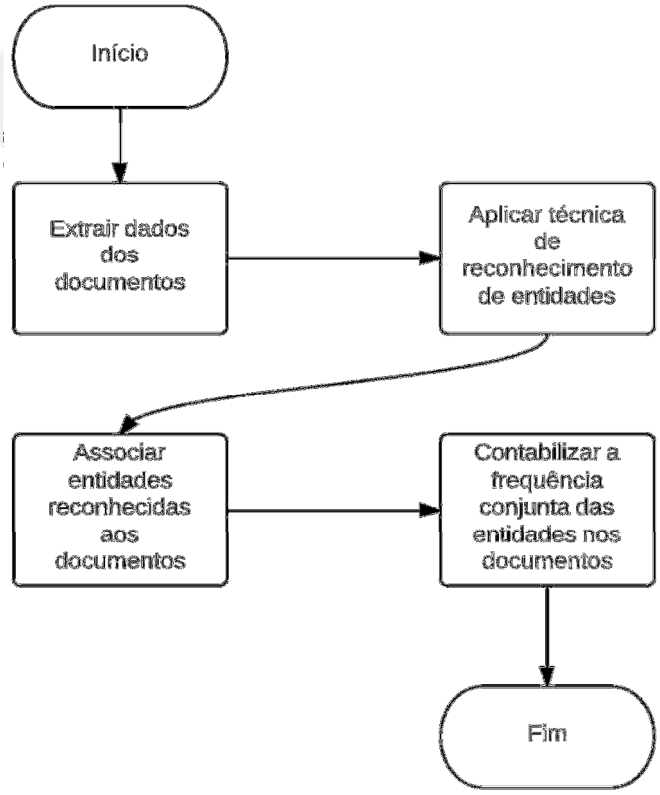
ISBN 978-85-61115-09-8

3 Experimento

O experimento para a descoberta de conhecimento é composto por dois macroprocessos: o macroprocesso de extração e transformação de dados e o processo de geração de visualização dos dados.

O macroprocesso de extração e transformação dos dados engloba desde as atividades de extração de dados a partir da fonte de informação até o armazenamento das entidades reconhecidas e correlacionadas na estrutura de dados definitiva. A Figura 1 apresenta os subprocessos que compõem essa parte da abordagem.

Figura 1 – Fluxograma do macroprocesso de transformação dos dados



Fonte: Elaboração dos autores (2014)

A sequência inicia-se com a recuperação dos dados a partir da fonte de informação. Neste passo, é realizado o tratamento necessário para obter as informações não estruturadas que serão utilizadas para análise no domínio. Estas informações são agrupadas de acordo com os documentos detentores das mesmas.

O processo seguinte é o reconhecimento de entidades a partir das informações dos documentos selecionadas no passo anterior. A técnica de reconhecimento de entidades adotada para extrair as entidades utiliza o aprendizado não-supervisionado como base de funcionamento. O processo não classifica as entidades por domínio, focando somente na descoberta de entidades para a execução do processo de correlação.

As informações não estruturadas são processadas pelo extrator, que aplica técnicas estatísticas sobre o *corpus* de análise. As palavras que compõem as informações são agrupadas de forma que sejam geradas todas as combinações possíveis de dois até cinco termos em sequência na frase. As combinações iniciadas ou terminadas por palavras de maior

ISBN 978-85-61115-09-8

uso (tais como preposições e artigos) são descartadas da análise. As combinações resultantes deste processo são denominadas entidades candidatas.

As entidades candidatas são repassadas para uma estrutura de dados responsável por contabilizar a frequência das combinações durante a análise de cada um dos resumos do objetivo. No final da análise, cada combinação que teve ao menos uma reincidência durante a análise é validada através de um dicionário de termos válidos no domínio em questão, de forma que palavras com erros ortográficos sejam descartadas do reconhecimento e que as palavras que não estejam relacionadas ao domínio sejam descartadas. Caso todos os termos que compõem a combinação ocorram no dicionário, a combinação recebe o estado de entidade reconhecida, sendo disponibilizada para utilização posterior no macroprocesso.

O terceiro processo compreende a associação dos termos descobertos no processo de reconhecimento de entidades. Cada um dos documentos extraídos no primeiro passo é analisado de forma que se possa verificar se este possui uma ou mais entidades reconhecidas. Em caso afirmativo, a entidade é associada ao documento.

Por fim, é realizada a contagem de ocorrências e coocorrências. Os documentos são analisados de forma que se possa obter a frequência individual de cada um dos termos e a frequência conjunta de todas as combinações únicas possíveis de dois termos associados (caso existam pelo menos dois termos associados ao documento).

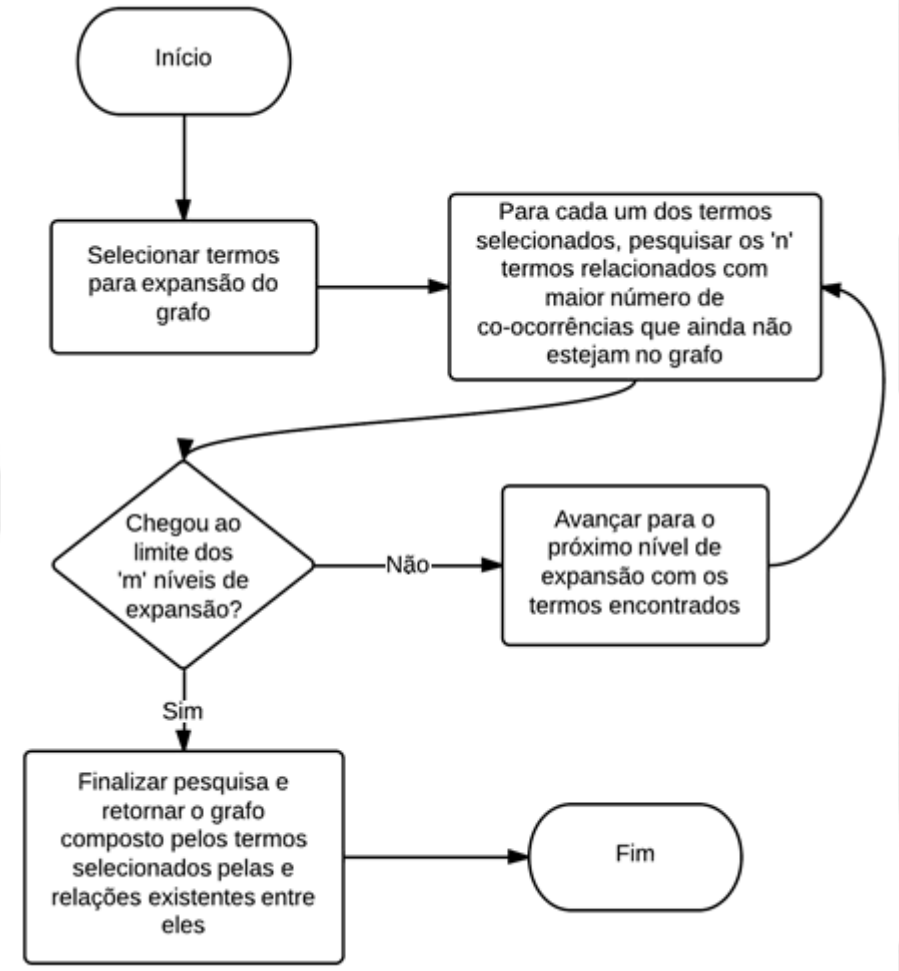
Esse macroprocesso realiza uso intensivo de processamento computacional e demanda tempo para que sua execução seja concluída, que por sua vez é proporcional ao número de documentos e ao número de entidades reconhecidas. No estudo de caso realizado na Plataforma Aquarius (o qual será apresentado na seção a seguir), o macroprocesso foi segmentado em dois aplicativos, executadas em *batch*[15]. O macroprocesso é executado novamente para cada atualização do banco de dados de convênios.

O macroprocesso de geração de visualização dos dados engloba o fluxo de ordens e regras de utilização da estrutura de dados definitiva para a montagem do mapa de conhecimento, que apresenta as informações processadas e sumarizadas. A Figura 2 demonstra essas atividades.

ISBN 978-85-61115-09-8

[15] Formato de execução de aplicativos no qual não há intervenção manual de um operador ou usuário, sendo que após o término das atividades programadas, o mesmo se encerra automaticamente.

Figura 2 – Fluxograma do macroprocesso de geração de visualização dos dados



Fonte: Elaboração dos autores (2014)

A sequência de processos inicia-se com a seleção dos termos para visualização na raiz da árvore. Caso o usuário informe um ou mais termos, estes são utilizados como termos raízes do gráfico. Se o usuário não informar algum termo, o processo selecionará as cinco entidades associadas aos convênios com maior frequência individual.

A partir dos termos raízes, inicia-se um ciclo de atividades, em que o número de ciclos ocorre de acordo com os dados disponíveis na estrutura e dos números máximos de níveis e de número de filhos por termo.

Para cada um dos termos do último nível processado, são pesquisados os cinco termos relacionados que ainda não estão representados no gráfico com maior quantidade de co-ocorrências. Estes termos (denominados termos-filho) são conectados com o termo do último nível processado (denominado termo-pai) através de arestas.

Caso o número de expansões ocorridas a partir da raiz não tenha ultrapassado o que tenha sido definido durante as regras de negócio, o mesmo processo é realizado para todos os termos-filhos encontrados neste nível. Desta forma, estes termos também se tornam termos-pai do próximo nível de expansão.

Ao atingir o limite de expansões, a árvore de resultados é transformada em um grafo que pode ser visualizado pelo usuário.

ISBN 978-85-61115-09-8

Na Plataforma Aquarius, este processo é realizado para cada requisição de montagem de mapa de tópicos. Foi disponibilizado um serviço interno na plataforma, que responde as solicitações com a execução do macroprocesso descrito de acordo com os parâmetros de configuração informados. O serviço utiliza a mesma estrutura de dados que foi criada e populada pelo macroprocesso de transformação de dados.

4 Estudo de caso: Plataforma Aquarius

A Plataforma Aquarius é uma plataforma integradora de dados abertos provenientes de diversos sistemas utilizados pela Corregedoria Geral da União (CGU) e pelo Ministério da Ciência, Tecnologia e Inovação (MCTI). O ambiente apresenta um conjunto de instrumentos gráficos e relatórios, agrupados como forma de responder a uma série de perguntas predefinidas nos temas de Dispêndios, Fundos Setoriais, Bolsas, Convênios e Produções Científica, Tecnológica e de Inovação. Os instrumentos de análise e relatórios apresentados nestas perguntas provem informações para a montagem de uma espécie de "sala de situação", que fornece informações que auxiliam o MCTI a decidir como e onde empenhar recursos para o aprimoramento dos resultados obtidos com a gestão. (MERCADANTE, 2012; SANTANA et al., 2012)

A abordagem apresentada nesse artigo foi aplicada na pergunta "Qual é a distribuição temática dos convênios?", que visa detectar quais são as principais palavras-chaves mencionadas no objeto dos convênios firmados pelo governo e sua correlação com outras palavras-chaves dentro do mesmo domínio. Como fonte de informação, foram utilizados os dados abertos disponibilizados pelo Sistema de Convênios do governo federal brasileiro (SICONV).

A partir do experimento, foi produzida uma base de dados com a frequência conjunta para cada combinação de duas entidades ocorrida no resumo do objetivo dos convênios. Além disso, foi elaborado um mecanismo de visualização dos dados *on-the-fly*[16], que permite a geração instantânea das informações presentes, abrangendo as seguintes regras:

a) o número máximo de termos-raízes da árvore é cinco;
b) caso exista somente um termo-raiz, este é considerado o centro do mapa de conhecimento e não contabiliza um nível de processamento no mapa;
c) caso exista mais de um termo-raiz, os mesmos são interligados a partir de um nodo central sem nenhum termo e disponibilizados como termos encontrados no primeiro nível de processamento;
d) o fator de expansão para cada termo (o número máximo de termos-filhos para cada termo) é cinco, sendo que o critério de seleção para exibição no mapa é a maior quantidade de coocorrências;
e) caso um termo selecionado para a expansão já exista no mapa, este é descartado para a expansão, sendo utilizado o próximo termo candidato;
f) o número máximo de níveis processados é três;

As regras foram adotadas com o intuito de disponibilizar uma quantidade significativa de termos relacionados sem haver a necessidade de se realizar a rolagem interna dentro do componente para verificar a existência de outros termos. A decisão sobre o número de níveis utilizados deve-se também a complexidade computacional existente no processo. Cada nível adicionado significa a realização de uma nova consulta para cada um dos termos disponíveis no nível anterior, caracterizando uma progressão exponencial de complexidade.

[16] Tipo de processamento de atividades computacionais realizado dinamicamente, iniciando-se a partir do momento no qual o resultado dessa atividade é requerido por um usuário.

ISBN 978-85-61115-09-8

O formato de resposta adotado representa a estrutura de um grafo através da descrição dos nodos e das arestas que o compõem. A disposição gráfica dos nodos e arestas ficou a cargo do componente gráfico adotado para exibição e disponibilizado pela plataforma.

Os resultados obtidos com a execução do algoritmo sumarizam as informações sobre os convênios em diferentes perspectivas. Neste trabalho serão apresentados dois mapas de tópicos gerados a partir dessa base de dados produzida, observando determinados critérios de geração.

A Figura 3 exibe o mapa de conhecimento gerado a partir das cinco entidades mais frequentes de todos os convênios assinados entre 2010 e 2012.

Figura 3 – Página dos instrumentos de análise para gerência estratégica da pergunta "Qual a distribuição temática dos projetos?" da Plataforma Aquarius, utilizando como critério de filtragem o triênio 2010-2012

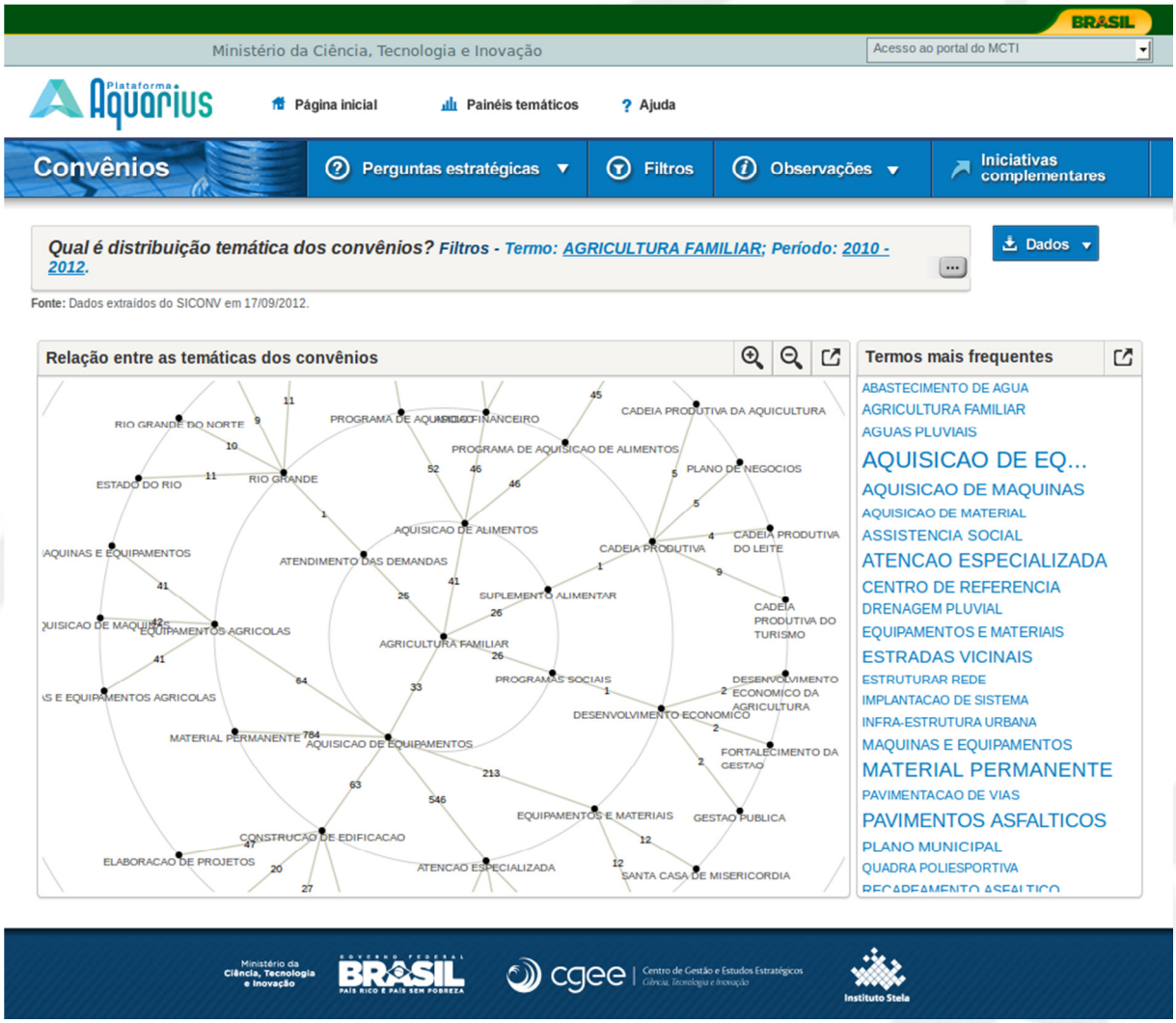
Fonte: Ministério da Ciência, Tecnologia e Inovação (2012)

A partir do mapa da Figura 3 denota-se que a maior parte dos convênios possui como temática a aquisição de equipamentos, que está relacionada à construção de edificações e a equipamentos agrícolas. Outra relação que pode ser identificada é a de que os convênios que possuem como objetivo a pavimentação asfáltica estão relacionados à pavimentação de vias urbanas e a drenagem de águas pluviais.

ISBN 978-85-61115-09-8

A Figura 4 exibe o mapa de conhecimento gerado a partir do termo “agricultura familiar”, considerando também todos os convênios assinados entre 2010 e 2012.

Figura 4 – Página dos instrumentos de gestão estratégica da pergunta “Qual a distribuição temática dos projetos?” da Plataforma Aquarius, utilizando como critérios de filtragem o termo raiz “agricultura familiar” e o triênio 2010-2012



Fonte: Ministério da Ciência, Tecnologia e Inovação (2012)

O mapa apresentado na Figura 4 apresenta outras relações para o termo “agricultura familiar”. Neste mapa, é possível perceber a existência de relação entre a agricultura familiar e os programas sociais (o que indica a existência de programas sociais dedicados à agricultura familiar), assim como a relação entre a agricultura familiar, a aquisição de alimentos e programas dedicados à aquisição de alimentos (o que indica a existência de programas para a utilização de alimentos provenientes da agricultura familiar).

Em ambos os cenários, o experimento conduzido possibilitou capturar o resumo do objetivo dos convênios e apresentá-lo em forma de palavras-chaves.

É possível gerar outros mapas de tópicos, utilizando faixas de triênios, programas ou estados específicos a partir da Plataforma Aquarius. Os mapas de tópicos podem ser gerados a partir da pergunta “Qual a distribuição temática dos convênios?”, disponível no painel de Convênios.

ISBN 978-85-61115-09-8

Considerações Finais

O experimento demonstrou a viabilidade de se sumarizar grandes quantidades de informação não estruturada e esquematizá-las em um formato de apresentação nos quais os principais tópicos e suas interligações são destacados. A abordagem em si é genérica, o que permite reaplicá-la a diferentes domínios, embora a qualidade da informação obtida esteja fortemente relacionada ao tipo de texto fornecido como entrada.

É possível realizar ajustes na abordagem, de forma a enquadrá-la melhor para determinado domínio. Citando como exemplo, o processo de reconhecimento de entidades pode adotar procedimentos baseados em clusterização ou uma lista prévia de verbetes a serem considerados válidos. O processo de geração de matrizes de correlação também pode ser refinado, considerando janelas de termos para que determinados termos sejam considerados correlacionados.

Tanto a técnica de reconhecimento de entidades quanto à de geração de matriz de correlação utilizam muitos recursos de hardware, dependendo do volume de documentos a serem interpretados e traduzidos para o modelo da matriz. O estudo de aplicação da abordagem realizado na Plataforma Aquarius demonstrou que o tempo para identificação de entidades e verificação de coocorrências entre elas levou algumas horas. No entanto, a aplicação de técnicas de processamento paralelo e de utilização de estruturas de dados otimizadas para o propósito da abordagem podem reduzir o tempo dispendido.

O complemento trazido por esse tipo de abordagem para um cenário onde somente dados estruturados são analisados e sumarizados pode ser bastante relevante para o processo de tomada de decisão. Dentre esse tipo de informação podem estar documentos organizacionais, opiniões de usuários sobre um determinado fato ou ainda campos textuais discursivos existentes em sistemas transacionais (como campos de observações ou justificativas, informados como texto pelo operador do sistema em determinadas rotinas que comportem esse tipo de informação).

Dentre as possibilidades de expansão existentes para a abordagem, são apresentadas abaixo algumas ideias que podem aprimorar os resultados obtidos:

a) classificação das entidades reconhecidas: os termos descobertos pela técnica de reconhecimento de entidades poderiam ser classificados de acordo com o programa ou com a faixa do valor pago pelo convênio;
b) utilização do coeficiente de correlação de Pearson: as ligações entre os termos poderiam apresentar também o peso que estas possuem dentro do conjunto de documentos;
c) autilização do coeficiente de associação entre termos: novas ligações poderiam ser estabelecidas através de associações indiretas (onde existe um termo intermediário em comum entre os termos a serem conectados), possibilitando assim a visualização de conexões entre termos que não ocorrem em conjunto, mas possuem uma conexão.

Referências

BRASIL. Lei nº 10.707, de 30 de julho de 2003. **Dispõe sobre as diretrizes para a elaboração da lei orçamentária de 2004 e dá outras providências**. Retirado em 6 de outubro de 2012, de: http://www.planalto.gov.br/ccivil_03/leis/2003/l10.707.htm.

BRASIL. Decreto nº 6.170, de 25 de julho de 2007. Dispõe sobre as normas relativas às transferências de recursos da União mediante convênios e contratos de repasse, e dá outras providências. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/decreto/ d6170.htm>. Acesso em: 29 jul. 2012.

ISBN 978-85-61115-09-8

BRASIL. **Convênios e outros repasses**. 3ª ed. Brasília: TCU, 2009. Disponível em: <http://portal2.tcu.gov.br/portal/pls/portal/docs/2053252.PDF>. Acesso em: 05 out. 2012.

BRASIL. **Licitações e contratos**. 4ª ed. Brasília: TCU, 2010. Disponível em: <http://portal2.tcu.gov.br/portal/pls/portal/docs/2057620.PDF>. Acesso em 29 set. 2012.

DICIONÁRIO Priberam da Língua Portuguesa [online]. 2008-2013. Lisboa: Priberam Informática, 2014. Corpus. Disponível em: <http://www.priberam.pt/DLPO/corpus>. Acesso em: 23 out. 2014.

GRIMES, S. **Unstructured data and the 80 percent rule**: Investigating the 80%. Washington D.C.: Bridgepoint. 2008. Disponível em: <http://www.clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551>. Acesso em: 28 jan. 2013.

KOZAREVA, Z.; FERRÁNDEZ, O.; MONTOYO, A. **Combining data-driven systems for improving Named Entity Recognition**. Data and Knowledge Engineering, 2007. Disponível em: <http://www.isi.edu/~kozareva/papers/nldb05cc.pdf>. Acesso em: 15 ago. 2012.

KRUPKA, G. R.; HAUSMAN., K. Isoquest, Inc: Description of the NetOwl(TM) extractor system as used for MUC-7. MESSAGE UNDERSTANDING CONFERENCE, 7., 1998, Virginia. Disponível em: <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/ proceedings/muc_7_proceedings/isoquest.pdf>. Acesso em: 19 ago. 2012.

MERCADANTE, A. **A Plataforma Aquarius, por Aloizio Mercadante**. Disponível em: <http://aquarius.mcti.gov.br/app/nova-governanca-publica/>. Acesso em: 14 out. 2012.

MIKHEEV, A.; MOENS, M.; GROVER, C. Named Entity recognition without gazetteers. EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 9., 1999, Stroudsburg. Disponível em: http://www.ltg.ed.ac.uk/ np/publications/ ltg/papers/Mikheev1999Named.pdf. Acesso em: 18 ago. 2012.

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO. **Plataforma Aquarius**. Disponível em: <http://aquariusp.mcti.gov.br/app/>. Acesso em: 15 out. 2012.

MINISTÉRIO DO PLANEJAMENTO, ORÇAMENTO E GESTÃO. **Boletim gerencial do SICONV n. 1. Brasília: Ministério do Planejamento, Orçamento e Gestão**. Brasília: TCU, 2009. Disponível em: <https://www.convenios.gov.br/portal/arquivos/ Boletim_Gerencial_SICONV_n01.pdf>. Acesso em: 5 out. 2012.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, v. 30, n. 1, p. 3-26, 2007. Disponível em: <http://nlp.cs.nyu.edu/sekine/papers/li07.pdf>. Acesso em: 19 ago. 2012.

PECK, R.; DEVORE, J. L. **Statistics: The Exploration and Analysis of Data**. 7. ed. Boston: Brooks/Cole, 2012.

SANG, E. F. T. K.; DE MEULDER, F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. CONFERENCE ON NATURAL LANGUAGE LEARNING, 7., 2003, Edmonton. Disponível em: <http://nlp.cs.swarthmore.edu/ ~richardw/papers/ tjongkimsang2003-conll.pdf>. Acesso em: 19 ago. 2012.

SANTANA, P. H. de A.; NEHME, C. C.; MOTA, J. P.; MACIEIRA, A.; MADRUGA, J. Transparência e modernização da gestão do Ministério de Ciência, Tecnologia e Inovação (MCTI) a partir da revisão e automação de processos no âmbito da Plataforma Aquarius. CONGRESSO CONSAD (CONSELHO NACIONAL DE SECRETÁRIOS DE ESTADO DA ADMINISTRAÇÃO) DE GESTÃO PÚBLICA, 7., 2012, Brasília. Disponível em: <http://www.escoladegoverno.rn.gov.br/contentproducao/aplicacao/searh_eg/imprensa/pdf/105.pdf>. Acesso em: 19 ago. 2012.

TABACHNICK, B. G.; FIDELL, L. S. **Using Multivariate Statistics**. 5. ed. Boston: Pearson, 2007.

TAN, H. **Knowledge Discovery and Data Mining**. 1ª ed. Berlin: Springer, 2012.

ISBN 978-85-61115-09-8

RELAÇÃO ENTRE LOD E MOOC MEDIADA POR OERs

Viviane Helena Kuntz
vkuntz@gmail.com

Luiz Antônio Moro Palazzo

luiz.palazzo@gmail.com

Vania Ribas Ulbricht
vrulbricht@gmail.com

Resumo
As tecnologias, práticas, abordagens e formatos emergentes estão se concentrando a partir do conceito *open*. Neste artigo ressalta-se as temáticas de Massive Open Online Course (MOOC) e Linked Open Data (LOD). Propõem-se, para tanto, um estudo exploratório que visa relacionar as temáticas. Nesse sentido, inicialmente buscou-se nas bases de dados identificar a literatura pertinente e em um segundo momento analisou-se as características, uso, aplicações para cada temática. Esta proposta apresenta, como resultado, exemplos de possíveis aplicações e futuras pesquisas sobre como relacionar MOOCs e LOD, mediada porOERs.

Palavras-chave: Massive Open Online Course (MOOC), Linked Open Data (LOD).Open Education Resource (OER)

Abstract
Technologies, practices, emerging approaches and formats are focusing from the open concept. This paper points out the themes of Massive Open Online Course (MOOC) and Linked Open Data (LOD). It is proposed, therefore, an exploratory study to relate the themes. Accordingly, we sought initially in databases to identify the relevant literature and in a second step we analyzed the characteristics, use, applications for each theme. This proposal has, as a result, examples of possible applications and future research on how to relate LOD and MOOCs mediated OERs.

Key Words:, Massive Open Online Course (MOOC), Linked Open Data (LOD).Open Education Resource (OER)

1 Introdução

As práticas, recursos, abordagens, formatos e sistemas abertos são realidades recorrentes estudadas em diversas áreas como: política, tecnológica, educacional etc. Na área educacional, sendo caracterizada como abordagem emergente tem-se os Massive Open Online Course (MOOC). Já na área tecnológica, com foco nesse artigo o Linked Open Data (LOD).

Na etapa inicial de revisão a busca prévia das temáticas apontou que a relação dessas temática open – MOOC e LOD - devem ser mediadas por Open Education Resource (OER). Para tanto, propõem dar início conceituando-as e compreendendo cada uma delas:

a) **Linked Open Data (LOD)**: introduzidas em 2006 por Tim Berners- Lee em sua arquitetura Web nota LinkedDados e tornaram-se conhecidos como os princípios Linked Data refere-se a um conjunto de melhores práticas para publicação e interligação de dados estruturados na web. Os LODs deve ser publicado em conformidade com os princípios destinados a facilitar as ligações entre os conjuntos

ISBN 978-85-61115-09-8

de dados, conjuntos de elementos e vocabulários valor. (RÍOS HILARIO; FERNÁNDEZ; CAMPO, 2013).

b) **Open Education Resource (OER):**veio pela primeira vez para usar em 2002 em uma conferência organizada pela UNESCO.Os participantes desse fórum definido OER como: " A provisão aberta de recursos educacionais, habilitados pelas tecnologias de informação e comunicação, para consulta, uso e adaptação por uma comunidade de usuários para fins não comerciais. (ULLRICH, 2008).

c) **Massive Open Online Course (MOOC):** passou a ser difundido a partir de 2008, quando Wiley, iniciou o desenvolvimento de recursos educacionais abertos, possibilitando a existência de a uma nova modalidade de cursos online. Estes cursos tem como característica permitir a participação de milhares de pessoas das diversas regiões do planeta, sem nenhum custo financeiro, favorecendo o uso por pessoas com dificuldades financeiras (CARVALHO, ET, 2013, p.205).

Parte-se dessa ordem lógica dos conceitos no texto LOD, OER e MOOCs, pois inicialmente tem-se uma estruturação especial de dados LOD), avançamos para um produto educacional (OER) e chegamos a uma aplicação que emprega este produto, no caso o MOOC.

Para tanto, propõem-se como estrutura o aprofundamento do termo LOD e MOOC e posteriormente a identificar a relação por meios dos OERs.

2. Linked Open Data (LOD)

A Web modificou radicalmente a forma de compartilhar conhecimento entre as pessoas, reduzindo as barreiras para publicar e acessar documentos como parte de um espaço global de informações. Isto foi possível devido a natureza aberta e extensível da Web, o que é considerado também um fator chave para tornar o seu crescimento exponencial. No entanto, apesar dos inegáveis benefícios produzidos, até recentemente os mesmos princípios que permitiram o florescimento da Web de documentos não tinham sido estendidos para uma Web de dados [BIZER, HEATH and BERNERS-LEE, 2009]. Segundo os mesmos autores, dados são tradicionalmente publicados na Web em forma bruta, em formatos como CSV ou XML ou ainda como tabelas HTML, sacrificando muito de sua estrutura e semântica. No entanto, nos últimos anos, a Web tem evoluído de um espaço global de informações baseado em documentos para outro, onde tanto documentos como dados estão ligados. Sustentando esta evolução há um conjunto de boas práticas para a publicação e conexão de dados estruturados na Web, conhecido como Linked Data. O W3C (World Wide Web Consortium) define linked data da seguinte maneira:

> Para tornar a Web Semântica ou Web de Dados uma realidade, é necessário interconectar um grande volume de dados disponíveis na Web em um formato padrão, acessível e manipulável. Além disso, o relacionamento entre esses dados precisa ser tornado explícito. Esta coleção de dados interrelacionados na Web pode ser denominada linked data e reside na própria essência da Web Semântica. (W3C - http://www,w3,org/standards/semanticweb/data/)

Pode-se assim pensar em linked data como uma forma de usar a Web para criar links tipados entre dados de diferentes fontes. Tecnicamente, a expressão se refere a dados publicados na Web de forma a serem "compreendidos por máquina". Sua semântica é explícita e estão representados de forma que tanto podem conectar-se e interagir com conjuntos de dados externos, quanto ser por sua vez conectados e demandados por estes.

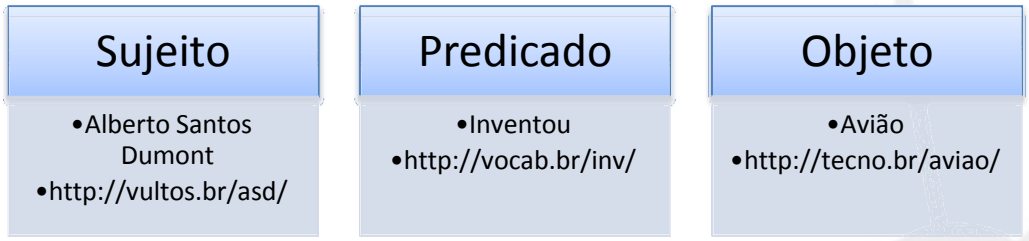
ISBN 978-85-61115-09-8

Tim Berners-Lee (2006) propôs um conjunto de regras para a publicação de dados na Web de forma que todos os dados assim publicados viessem a se tornar parte de um único espaço de dados global:

a) Usar URIs para dar nomes às coisas.
b) Usar URIs expressas em HTTP, de forma que possam ser pesquisadas.
c) Adicionar informação útil às URIs por meio de padrões (RDF, SPARQL).
d) Incluir links para outras URIs de modo que outras coisas possam ser pesquisadas.

Estas regras se tornaram informalmente os princípios do linked data e oferecem uma receita básica para publicar e conectar dados usando a infraestrutura da Web. Essenciais aqui são duas tecnologias empregadas na Web: URI (Uniform Resource Identifier) [BERNERS-LEE, 2005] e HTTP (Hyper Text Transfer Protocol) [FIELDING et al, 1999]. As URIs e o protocolo HTTP são complementados por uma tecnologia crítica para a Web de Dados - o padrão RDF (Resource Description Framework), uma linguagem simples para descrever triplas do tipo "sujeito-predicado-objeto" que, em conjunto, representam estruturas em forma de um multigrafo semântico, que pode ser navegado, pesquisado e empregar técnicas de inferência para responder perguntas formuladas por humanos ou por máquinas. RDF possui um nível de esquema (RDFS) que define os conceitos, relacionamentos e os aspectos estruturais da representação. As instâncias das entidades formalizadas no esquema são representadas no nível objeto (RDF propriamente dito). Um exemplo de uma tripla RDF típica é apresentado na figura 1.

Figura 1 - Uma Tripla RDF



Fonte: Elaborada pelos autores

Um conceito RDF normalmente é representado sintaticamente por meio de uma linguagem de markup baseada em XML. Assim um documento RDF representa fundamentalmente um conjunto de triplas sujeito-predicado-objeto (ou, no nível do esquema, conceito-relacionamento-conceito) expresso na sintaxe XML. Como as entidades descritas no documento são todas identificadas por URIs e associadas via endereços HTTP a objetos que se encontram alhures na Web de Dados, o que se tem na realidade é um objeto de conhecimento distribuído, que pode ser "inteligentemente" interpretado e processado por máquinas. O objetivo final do linked data é permitir o uso da Web como uma única base de dados global. Conforme advertem Bizer et al (2009), a realização desta visão é capaz de beneficiar inúmeras áreas mas, por outro lado, podem agravar os problemas encontrados em outras. Uma dessas áreas problemáticas é a que surge das oportunidades de violação de privacidade que surge da integração de dados de diferentes fontes. Aqui é importante considerar o trabalho de Weitzner (2007) sobre o paradoxo da privacidade na Web de dados.

3. Massive Open Online Course (MOOC)

A sigla MOOC (Massive Open Online Curso) tem suas raízes na educação a distância,que é fundamentada no conceito de Open Education Resource (OER), aliada a necessidade de massificar o ensino.

ISBN 978-85-61115-09-8

Experiências dispersas surgiram por meio das novas Tecnologias da Informação e da Comunicação até chegarmos às iniciativas atuais. (NICOLAU, NICOLAU, 2014, p.1)

Miguel (2012) apresenta um conceito elucidativo deste tipo de cursos:

> Os MOOCs representam experiências de aprendizagem inovadoras baseadas nas tecnologias de informação e comunicação, em plataformas *web* 2.0 e redes sociais. A participação em um MOOC é aberta para qualquer interessado e envolve grande quantidade de material didático, o conteúdo é produzido pelos alunos em diversos canais de expressão, como os *blogs*, fóruns, compartilhamento de imagens, vídeos, áudio entre outros recursos, o conhecimento é construído a partir do envolvimento direto dos alunos em função de interesses em comum. Não existe uma pedagogia propriamente dita movendo a interação cooperativa, porém surgem metodologias e novas dinâmicas de socialização que favorecem a troca de informações e, por conseguinte, de conhecimento. (MIGUEL, 2012, p. 123).

Contrapondo os autores já citados para Carvalho, *et al.* (2013) os MOOCs tratam-se de uma modalidade de ensino que se utiliza dos recursos e ferramentas da educação a distância (EaD). Para o autor esses cursos, frequentemente utilizam os mesmos elementos de:

a) ensino a distância, como conteúdos em formato de textos, imagens, áudios e vídeos, além de fóruns de discussão para favorecer a aprendizagem e a interação entre os participantes;
b) flexibilidade (o aluno podem escolher os melhores horários par estudar);
c) formação de grupos (alunos interessados podem se organizar para estudar).

No entanto, tem características de ausência de pré-requisitos e de emissão de certificados. (CARVALHO, *ET. AL.*, 2013). Gaebel (2013, p.3) afirma que até o momento, as instituições de ensino superior que oferecem MOOCs afirmaram que não concedem créditos, mas apenas uma declaração de cumprimento do curso. Mas, como muitas questões relacionadas MOOCs, esta prática tende a desaparecer no futuro.

Ainda como características, Stuchlikova e Kosa (2013) e Gupta e Sambyal (2013, p. 312-313) salientam a flexibilidade, acessibilidade, imersão direta e engajamento dentro do tema, gratuidade ou baixo custo, alcance do público alvo, ausência de pré requisitos, e variedade de ferramentas.

No sentido de estrutura do MOOC Nicolau e Nicolau (2014) afirmam que é ofertado a partir do AVA e de ferramentas da Web 2.0, não exigem pré-requisitos, podendo oferecer certificados de participação (NICOLAU, NICOLAU, 2014, p.1).

Além da definição, também faz-se importante a distinção entre os dois tipos de MOOCs, chamados cMOOC e xMOOC. Gonçalves (2013) demonstra as diferenças na autonomia, conteúdo e professor, conforme quadro 1.

Quadro 1 – Diferenças entre cMOOC e xMOOC

	cMOOC	***XMOOC***
Autonomia do Participante	Total: O participante tem que gerar e procurar informação externa (além do material disponível).	Parcial: O participante é conduzido pelo conteúdo do professor, mas também pode contribuir com conteúdo externos.
Conteúdo do Curso	Descentralizado: Enriquecido por conteúdos externos e pela partilha	Centralizado: Conteúdo principal fornecido pelo professor. Os

ISBN 978-85-61115-09-8

	de informações entre os vários participantes.	participantes podem trocar ideias na própria plataforma do curso.
Professor	Direciona apenas algumas informações partilhadas pelos participantes.	Centralizado: Conteúdo principal fornecido pelo professor. Os participantes podem trocar ideias na própria plataforma do curso.

Fonte: Gonçalves (2013, p.32)

Carvalho (2013) acrescenta que:

a) xMOOC: por suas características e organização mais rígidas, conferem maior ênfase aos conteúdos e ao ambiente adotado para a interface entre os usuários do curso. Portanto, apresenta uma abordagem hierarquizada, sem grande margem para estratégias do tipo tentativa-erro-reflexão ou para uma ação de caráter autoral do aluno em rede.
b) cMOOCs: é inspirada no modelo implementado e difundido desde 2008 por Siemens e Downes a partir do curso intitulado Connectivism and Connective Knowledge (Conectivismo e Conhecimento Conectador na tradução livre para o português). Desta experiência deriva o modelo fundamentado no Conectivismo, denominado cMOOC.

4 Procedimentos Metodológicos

Estapesquisa pode ser caracterizada como exploratória com abordagem qualitativa. Inicialmente realizou-se uma pesquisa bibliográfica, a partir dos termos "massive open online course" + "Linked Open Data". A pesquisa foi realizada nas bases de dados Scopus e no Google Acadêmico.

Com os artigos encontrados, vida quadro 2, buscou-se identificar características, uso, aplicações para cada temática.

Quadro 2 - ocorrências da busca

Autores	**Título**	**Ano**	**Periódico-Evento**	**Base**
PIETRA, N. , CHICAIZA, J., LÓPEZ, J., TOVAR CARO, E.	Supporting openness of MOOCs contents through of an OER and OCW framework based on Linked Data technologies	2014	*IEEE Global Engineering Education Conference, EDUCON*	*Scopus*
MIKRPYANNIDIS, A., DOMINGUE, J	Interactive learning resources and linked data for online scientific experimentation	2013	*WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*	*Scopus*

ISBN 978-85-61115-09-8

Autores	Título	Ano	Periódico-Evento	Base
TOMÁS, C. dos R.	Web semântica e personalização: repercussões da interação semântica com recursos educacionais abertos na identidade virtual do estudante e nos ambientes de aprendizagem online.	2013	*Dissertação de Mestrado*	*Google Scholar*
PIETRA, N. , CHICAIZA, J., LÓPEZ, J., TOVAR CARO, E	An Architecture based on Linked Data technologies for the Integration and reuse of OER in MOOCs Context.	2014	*Open Praxis*	*Google Scholar*

Visa-se, portanto, com essas referências identificar as relações e aplicações entre LOD e MOOC.

4 Relações da temáticas LOD e MOOC

Devido à representação em RDF, que emprega anotação semântica dos componentes, os objetos em linked data contém sua própria interpretação tanto em termos estruturais quanto em termos da semântica das entidades componentes. Isto os torna altamente compartilháveis, capazes de, uma vez publicados na Web, ser utilizados para consultas e produção de inferências. Esta capacidade faz com que os objetos em Link Open Data (LOD) possam ser facilmente vistos como objetos de conhecimento e empregados como Open Education Resource (OER) podem ser consultados por meio da linguagem SPARQL (Sparql Protocol and RDF Query Language). Isto possibilita a construção de interfaces gráficas que permitem a navegação e auto aprendizagem. Para Gary Matkin (2013), segundo a lógica evolutiva dos processos educacionais, MOOCs deverão ser em breve substituídos por OERs, na direção de um aprendizado mais autônomo e um currículo cada vez mais personalizado na graduação.

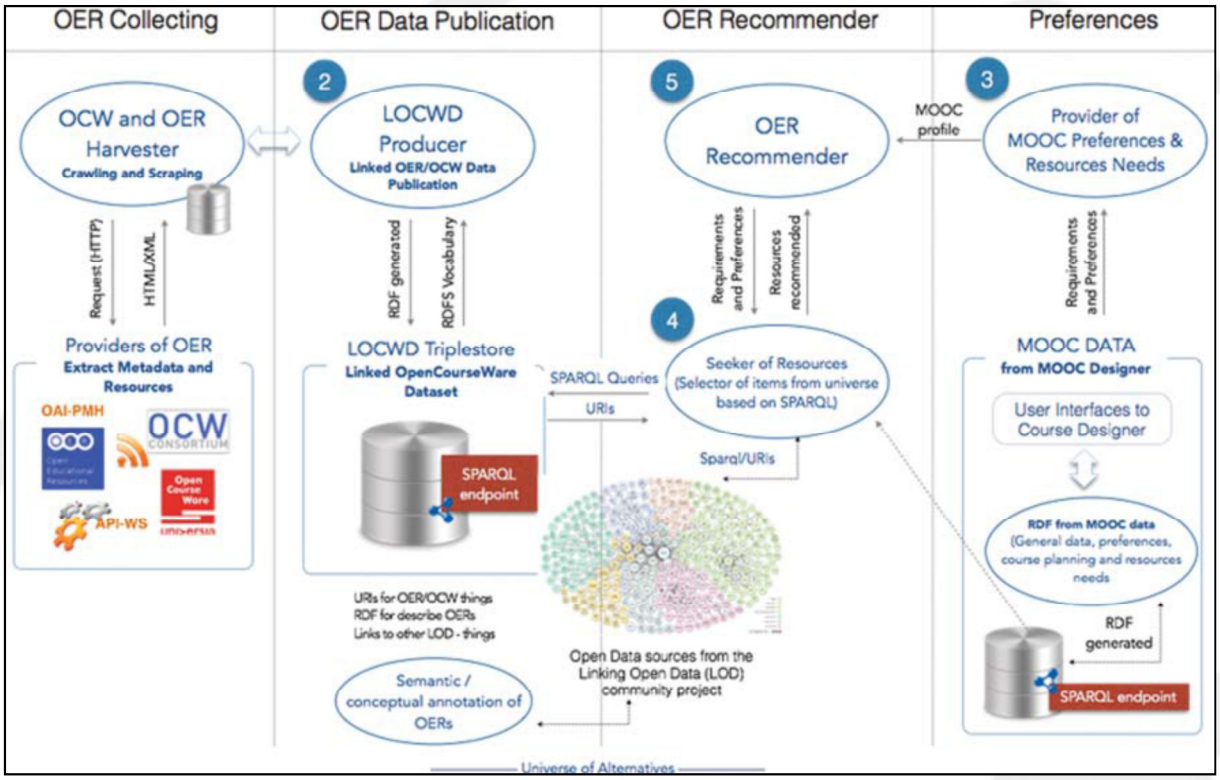
No entanto, esta opinião não é compartilhada por todos os pesquisadores da área. Segundo Piedra et al, (2014) MOOCs originaram-se da organização de OERs previamente existentes em contextos variados para a produção de unidades educacionais capazes de conferir uma determinada competência a um grande número de interessados. É fácil prever a automatização cada vez maior dos MOOCs, na medida em que, integrados ao espaço global de dados e informações, sejam capazes de oferecer currículos personalizados - e de excelente qualidade - inclusive para graus acadêmicos. A verdade entretanto é que o tema ainda é controverso e desperta o interesse de um grande número de pesquisadores das mais diversas áreas. Supostamente, há um ponto ideal de integração entre MOOCs e OERs, uma vez que estes podem ser vistos como matéria prima para desenvolver aqueles. É urgente portanto estabelecer critérios e padrões de design que permitam não apenas a utilização, mas também a criação de OERs por MOOCs.

ISBN 978-85-61115-09-8

5 Aplicações de Link Open Data e MOOC mediante OERs

Das ocorrências pesquisadas Piedra, N.; et al. (2014) propõem uma arquitetura baseada em tecnologias Linked Data para a Integração e reutilização dos REA em MOOCs, e em um outro artigo com os mesmos autores com objetivo deauxiliar MOOCs através de um quadro OER e OCW baseado em tecnologias Linked Data. Em ambos tem-se a arquitetura proposta esquematizada vista na figura 2.

Figura 2 - arquitetura comreutilização dos REA em MOOCs



Fonte: PIEDRA, N.; et al.

a. **Componente 1 (Coleção OER):** Identificar e selecionar os repositórios de REA, em seguida, extrair recursos de metadados e educacionais com Licenças Abertas
b. **Componente 2 (Publicação de dados OER):** Desenvolvimento e fornecimento de dados sobre os recursos educacionais abertos como Linked Data.
c. **Componente 3 (fornecer perfis MOOC**): extrair dados de um perfil de curso que serve como uma visão de filtro para todo o universo.
d. **Componente 4 (agrupar recursos):** unir as funcionalidades do candidato a recomendação e perfil no curso.
e. **Componente 5 (recomendação OER):** usar a arquitetura.

Para Mikrpyannidis, A., Domingue J. (2013), tem-se como proposta o FORGE, que trata-se de uma iniciativa europeia para aprendizagem online usando Future InternetResearch and Experimentation (FIRE). O FORGE é um passo para transformar FIRE em uma plataforma educacional (linha plataformas educacionais como o iTunes U, bem como em

ISBN 978-85-61115-09-8

MOOCs) por meio de Linked Data. Isso visa beneficiar alunos e educadores, dando-lhes o acesso a instalações de classe mundial a fim de realizar experimentos sobre, por exemplo, novos protocolos de Internet.

Tomás (2013) disserta sobre MOOC apoiados por tecnologia semântica, em que com o" auxílio da websemântica, a possibilidade da construção de ambientes de aprendizagem cada vez maispersonalizados, singulares e contextualizados (sem perder a diversidade da dimensão social). Para o autor, ainda:

> será possível, com o auxílio de mecanismos de inteligência artificial, ser introduzidos, por exemplo, perfis de questões e dúvidas, compatíveis com os diferentes indivíduos, através de uma automatização dos processos que garanta um alcance e uma acessibilidade global ao ensino, mas agora de forma personalizada. Esta possibilidade será viabilizada a partir dos dados que cada um vai colocando na web, a partir das múltiplas utilizações que faz digitalmente e que, ligados (Linked Data), potenciarão, cada vez mais, a personalização.

Nesse sentido aquilo que a dimensão semântica da Web possibilita, de acordo coma investigação efetuada, pode resumir-se em(Piedra et al., 2012, p. 30):

> 1) capacidade para auxiliar na integraçãodo trabalho disperso na rede de instituições produtoras de conteúdos (dados, para atecnologia semântica);
>
> 2) Os dados ligados (Linked) auxiliam:
> a) na descoberta dedados;
> b) na fiável reutilização de dados;
> c) no fornecimento melhorado em relação àsfontes de proveniência;
> d) na facilitação do processamento automatizado quer em relação à flexibilidade das mudanças na apresentação, quer na redução da ambiguidade

"Abertura e personalização são os ingredientes que potenciam a existência de REAs, MOOC e outras realidades educacionais ou novas modalidadesde ensino versus educação que, num mundo em mutação, possibilitam vários caminhospara a humanidade." (TOMÁS, 2013).

6 Considerações Finais

Com a elaboração desse artigo foi possível analisar os conceitos e características de Massive Open Online Course e Linked Open Data para assim, identificar as relações e aplicações entre essas temáticas.

As temáticas open estão ganhando espaço nas pesquisas acadêmicas, no entanto, ainda quanto se trata de relacionar abordagens, práticas e formatos o número de ocorrências reduz consideravelmente. Realidade essa comprovada quando se buscaram as temáticas "Linked Open Data" e "Massive Open Online Course" nas bases Scopus e no Google Acadêmico obtendo respectivamente duas e seis ocorrências, totalizando oito, porém em que quatro verificou-se a real relação.

As relações e aplicações das temáticas supracitadas são possível quando se tem a mediação do Open Educational Resource (OERs).

Verifica-se como futuros desdobramentos ampliação dos estudos de casos das aplicações propostas nas ocorrências obtidas.

ISBN 978-85-61115-09-8

Referências

BERNERS-LEE, Tim et al. (2005). Uniform Resource Identifier (URI): Generic Syntax. Request for Comments: 3986. Retrieved June 14, 2014, http://tools.ietf.org/html/rfc3986

BERNERS-LEE, Tim. (2006). Linked Data - Design Issues. Retrieved July 23, 2014 http://www.w3.org/DesignIssues/LinkedData.html

BERNERS-LEE, Tim et. al. (2006), Tabulator: Exploring and Analyzing Linked Data on the Semantic Web. Procedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06).

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim: (2009) Linked Data - The Story So Far. In Heath, T.; Hepp, M.; Bizer, C. (eds) International Journal on Semantic Web and information Systems; Special issue on Linked Data.

FIELDING, R. et al. (1999). Hypertext Transfer Protocol -- HTTP/1.1. Request for Comments: 2616. Retrieved June 14, 2009.

GONÇALVES, Bruno M. F. **MOOC e b-Learning**: uma proposta para o mestrado em TIC na Educação e Formação do Instituto Politécnico de Bragança. Dissertação. Instituto Politécnico de Bragança 2013.

MATKIN, Gary (2013): Open Educational Resources in the Post MOOC Era**. eLearn Magazine**. Retrieved June 2014, http://elearnmag.acm.org/archive.cfm?aid=2460460

PIEDRA, Nelson et al. Supporting openness of MOOCs contents through of an OER and OCW framework based on Linked Data technologies. In: **Global Engineering Education Conference (EDUCON)**, 2014 IEEE. IEEE, 2014. p. 1112-1117.

PIEDRA, Nelson et al. An Architecture based on Linked Data technologies for the Integration and reuse of OER in MOOCs Context. **Open Praxis**, v. 6, n. 2, p. 171-187, 2014.

RÍOS HILARIO, Ana B.; FERNÁNDEZ, Tránsito Ferreras; CAMPO, Diego Martín. Linked open bibliographic data. In: **Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality**. ACM, 2013. p. 333-337.

ULLRICH, Carsten et al. Why web 2.0 is good for learning and for research: principles and prototypes. In: **Proceedings of the 17th international conference on World Wide Web**. ACM, 2008. p. 705-714.

WEITZNER, D. (2007): Beyond Secrecy: New Privacy Protection Strategies for Open Information Spaces. IEEE Internet Computing, 11(5):94-96.

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

QUALISBRASIL: DISPONIBILIZANDO DADOS VIA *LINKED OPEN DATA* PARA ESTUDOS CIENTOMÉTRICOS

Sandro Rautenberg
srautenberg@unicentro.br

Edgard Marx
marx@informatik.uni-leipzig.de

Sören Auer
auer@cs.uni-bonn.de

Axel-C. Ngonga Ngomo
ngonga@informatik.uni-leipzig.de

Jens Lehmann
lehmann@informatik.uni-leipzig.de

Resumo
A obtenção de dados para estudos cientométricos é uma tarefa complexa, devido aos desafios em coletar, organizar e relacionar dados pertinentes. Tais desafios se dão porque a maioria dos dados está distribuída em várias fontes e é apresentada em formatos não estruturados ou proprietários na Internet. Para contornar essas dificuldades, tem-se o aporte do *Linked Open Data*. Este está fundamentado no conjunto de melhores práticas de *Linked Data* e em licenças abertas de utilização de dados. Em suma, isso permite que dados sejam consumidos ou reutilizados sem restrições, por diversos sistemas e contextos, de maneira facilitada. Com base no ciclo de vida *LOD2 Stack* e alguns estudos de caso, este trabalho é um relato da experiência na disponibilização do histórico dos dados do índice Qualis no formato *Linked Data*, permitindo sua utilização e interligação sem restrições em diversas aplicações.

Palavras-chave: *Linked Open Data*. *LOD2 Stack*. Qualis. Estudos Cientométricos.

Abstract
Obtaining data for Scientometric Studies is a complex task due to the challenges in collecting, organizing and relating relevant data. Such challenges occur because most of the data is distributed across multiple sources in the Internet and is presented in proprietary or unstructured formats. To overcome these difficulties, the Linked Open Data is used as technical support. It is based on the Linked Data practices, using open licenses to organize, connect and publish data. In a nutshell, it enables data to be consumed/reused without restrictions, in an easier way by different systems in several contexts. Based on the *LOD2 Stack* lifecycle and some use cases, this paper reports the experience in providing the historical data of Qualis Index to be used and interlink without restriction in many applications.

Key Words: Linked Open Data. *LOD2 Stack*. Qualis. Scientometric Studies.

ISBN 978-85-61115-09-8

Introdução

Este trabalho tem como foco um relato de experiência da publicação do índice Qualis em *Linked Data*, tendo como base a relação entre dois construtos para organizar, formalizar e compartilhar dados e/ou informações sobre a pesquisa científica: a Cientometria e o *Linked Open Data*.

Segundo o dicionário de Biblioteconomia e Arquivologia, a Cientometria é a "disciplina que tem por objetivo medir as atividades de pesquisa científica e tecnológica mediante insumos (mão-de-obra, investimentos) e produtos (equipamentos, produtos, publicações)" (CUNHA e CAVALCANTI, 2008). Numa visão geral, trabalhos desta natureza organizam e cruzam dados e informações sobre comunicações científicas (autores, artigos, patentes, grupos e institutos de pesquisa, entre outros) para nortear políticas de ciência e tecnologia, realizar a gestão do conhecimento em instituições de pesquisa, ou melhor conhecer um objeto de pesquisa. Como exemplos de estudos cientométricos encontrados na literatura, têm-se: a avaliação da pesquisa global sobre a educação: uma abordagem cientométrica (KONUR, 2012); perfil da produção científica brasileira sobre a gripe pandêmica de influenza A/H1N1 (LUCHS, 2012); a produção brasileira em Ciência da Informação no exterior como reflexo de institucionalização científica (ARBOIT et al.,2011); e co-autoria em artigos e patentes: um estudo da interação entre a produção científica e tecnológica (MOURA e CAREGNATO, 2011).

Contudo, na Cientometria, obter informações sobre o desenvolvimento da ciência para subsidiar ações/estudos pode ser uma tarefa complexa (SANTOS e KOBASHI, 2009). Envolve conhecimentos diversos, em especial de computação, visto que existem desafios na coleta, na organização e no relacionamento de dados pertinentes. Assume-se que tais desafios existem porque os dados são provenientes da Internet e estão distribuídos em várias fontes, apresentados em formatos distintos ou proprietários, dificultando a manipulação.

Toma-se como exemplo duas fontes de dados amplamente difundidas entre pesquisadores e disponibilizadas na Internet: a Plataforma Lattes (LATTES, 2013) e Sistema WebQualis (WEBQUALIS, 2013). Na Plataforma Lattes se encontram dados sobre as comunicações científicas produzidas por pesquisadores. Já no Sistema WebQualis estão estratificados os índices de qualidade de algumas dessas comunicações. Interligando estas fontes, é possível quantificar e qualificar algumas medidas da produção científica.

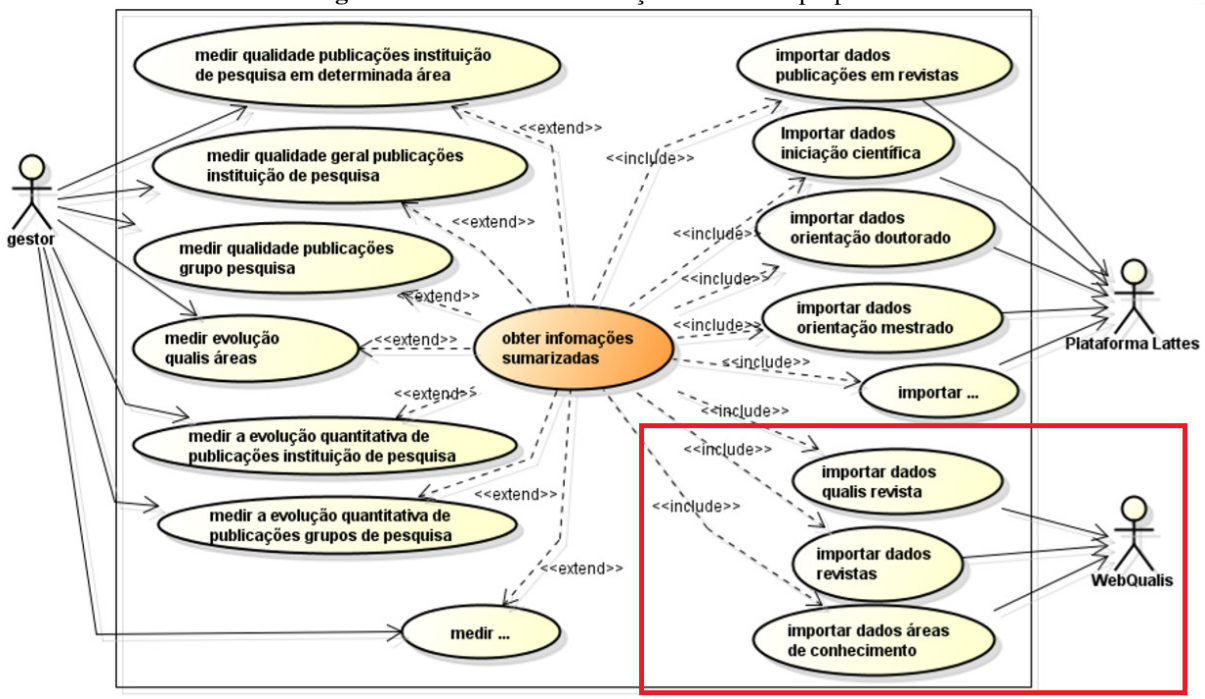
Mas, aferir informações a partir das fontes anteriormente citadas é uma tarefa custosa, visto que os dados são disponibilizados em formatos diferentes (os currículos Lattes se encontram em páginas de Internet e o índice Qualis dos periódicos estão no formato pdf), dificultando a organização, a extração, o cruzamento e a exploração de dados/informações.

Para contornar tais dificuldades, como abordagem da Web Semântica, tem-se o aporte do *Linked Open Data* (AKSW, 2013). Este se baseia no *Linked Data*, um conjunto de melhores práticas para organizar, publicar e conectar dados na *web* (LINKED DATA, 2012a). Adicionalmente, no *Linked Open Data*, os dados tratados são publicados sob licenças abertas, o que possibilita reutilização destes sem restrições, em diversos contextos e aplicações.

Diante dessa potencialidade de reutilização, a justificativa do trabalho se baseia no alinhamento das tecnologias de *Linked Open Data* para o tratamento, a organização, o cruzamento e a exploração de dados aderentes aos estudos do campo da Cientometria. Como perspectiva, pretende-se desenvolver um "Modelo Tecnológico ao Compartilhamento de Dados para Estudos Cientométricos baseado em *Linked Open Data*", conforme representado na Figura 1.

ISBN 978-85-61115-09-8

Figura 1 – Cenários de utilização do modelo proposto



Fonte: elaborado pelos autores.

Em face da complexidade do desenvolvimento deste modelo, este relato de experiência é delimitado à área em destaque na Figura 1. Para tanto, é traçado como objetivo o tratamento dos dados oriundos do Sistema WebQualis de acordo com os preceitos do *Linked Open Data*. Como resultado, tem-se a disponibilização de dados históricos do índice Qualis (2005-2013), para serem consumidos no desenvolvimento do modelo proposto e em demais pesquisas no campo da Ciência da Informação.

A fim de melhor discutir as experiências adquiridas ao publicar os dados, este relato aborda: i) a apresentação dos materiais e métodos utilizados no processo; ii) a discussão do ações realizadas para publicação; iii) a verificação da publicação frente alguns estudos de caso de consumo e vinculação de dados; e iv) a discussão dos resultados e a apresentação dos trabalhos futuros, seguidas pelos agradecimentos e referências bibliográficas.

1 Materiais e Métodos

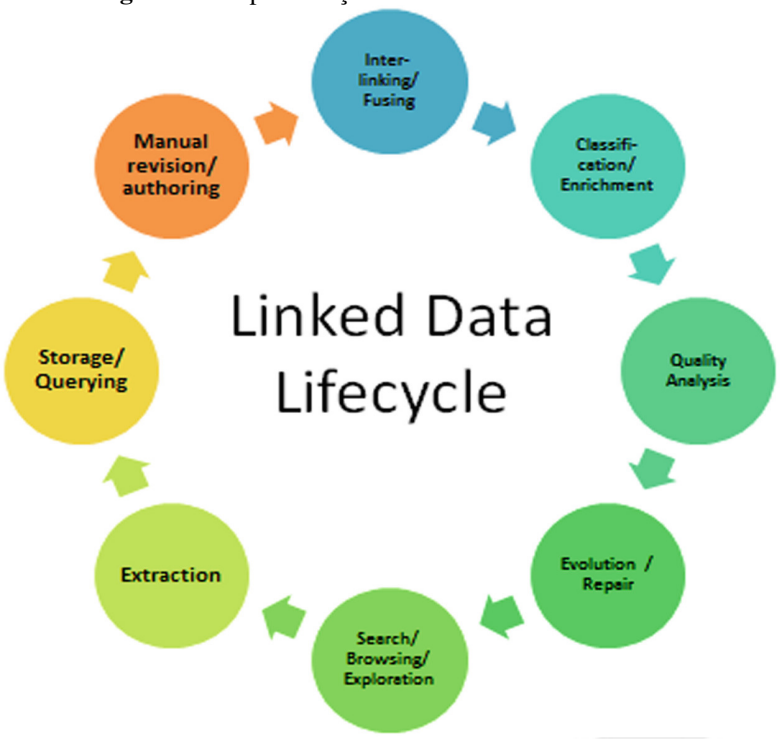
Como procedimento metodológico para publicação do índice Qualis é empregado o ciclo de vida de *Linked Open Data* baseado nas práticas do projeto LOD2 - *Creating knowledge out of Interlinked Data* (LOD2, 2014). Tal projeto é um empreendimento conjunto de importantes grupos de pesquisa que estão na vanguarda da evolução dos procedimentos metodológicos e das ferramentas computacionais voltadas ao *Linked Open Data*. Este ciclo de vida é denominado *LOD2 Stack* e ilustrado na Figura 2. Ele difunde um conjunto de oito estágios e suas ferramentas, empregados conforme a necessidade (AUER et al., 2012; AUER et al., 2013; AUER, 2014):

1. ***Extraction* –** mapear dados não estruturados, estruturados em diferentes formatos, ou provenientes de sistemas legados para um modelo de dados RDF (acrônimo de *Resourse Description Framework*).
2. ***Storage* –** utilizar soluções computacionais para armazenamento e recuperação de dados no padrão RDF.

ISBN 978-85-61115-09-8

3. ***Authoring* –** criar/editorar bases de conhecimento semanticamente enriquecidas, com o uso de tecnologias de *Semantic Wiki*, primando por aspectos como socialização, distribuição e colaboração.
4. ***Interlinking* –** empregar ferramentas para interligar dados de diversas fontes, ampliando os contextos para recuperação de dados/informações.
5. ***Classification* –** utilizar ontologias de alto nível para classificação/representação dos dados, melhorando a precisão em atividades de recuperação, integração e fusão de dados.
6. ***Quality* -** tratar pontualmente os aspectos de integridade, precisão, consistência e validade de dados; e de forma geral, verificar os quesitos de consistência, concisão, compreensibilidade, disponibilidade e proveniência dos modelos de dados.
7. ***Evolution/Repair* –** corrigir não conformidades causadas por inconsistências encontradas nos dados disponibilizados ou no modelo de representação perante requisitos estabelecidos.
8. ***Search/Browsing/Exploration* –** empregar soluções computacionais que facilitem as consulta, navegação e/ou exploração de dados, de acordo os objetivos do usuário.

Figura 2 – Representação do ciclo de vida *LOD2 Stack*



Fonte: AUER, 2014.

A Tabela 1 enumera as ferramentas empregadas no *LOD2 Stack* a cada estágio. Destaca-se que as ferramentas computacionais enumeradas são de livre utilização.

ISBN 978-85-61115-09-8

Tabela 1 – Ferramentas computacionais sugeridas a cada estágio

Estágio	Ferramentas
Extraction	OpenLink Virtuoso Sponger, DBpedia Spotlight, Dbpedia Spotlight UI, PoolParty, D2R, R2R, Apache Stanbol e CSVImport
Storage/Querying	OpenLink Virtuoso RDF Store, SparQLed, Sparqlify e SIREn
Manual Revision/Authoring	OntoWiki, RDFAuthor e PoolParty
Interlink/Fusion	LIMES, Silk, LACT Silk e Sieve
Classification/Enrichment	DL-Learner
Quality Analysis	ORE
Evolution/Repair	Google LODrefine
Search/Browsing/Exploration	Sigma, Spatial Semantic Browser, CubeViz e Facete

Fonte: LOD2 (2014).

Outras ferramentas livres não relacionadas na Tabela 1 também são utilizadas. A exemplo, com o Sistema Gerenciador de Banco de Dados Mysql (MYSQL, 2014), se organizou os dados, permitindo a geração de arquivos no formato CSV (Comma-Separated Values). E com o Script Lattes (MENA-CHALCO e CESAR JUNIOR, 2009) se extraiu os dados oriundos dos Currículos Lattes, de forma semi-automática, para a realização de estudos de caso.

Como materiais para publicação, além do vocabulário RDF e RDFS, são utilizados:

1. a tabela da classificação do índice Qualis (formato de planilha eletrônica - XLS) acessada no ano de 2007 e denominada neste trabalho de Qualis_2007;
2. a tabela da classificação do índice Qualis (formato Portable Document File - PDF) disponibilizada no ano de 2009 e denominada neste trabalho de Qualis_2009;
3. a tabela da classificação do índice Qualis (formato Portable Document File - PDF) acessada no ano de 2013 e denominada neste trabalho de Qualis_2013; e
4. o metadados Dublin Core (DCMI, 2014) para organização e representação dos dados em vocabulário amplamente discutido na Ciência da Informação.

Além disso, destaca-se a utilização do vocabulário SCOVO – *Statistical Core Vocabulary* (HAUSENBLAS et al., 2009) - na publicação de dados estatísticos num estudo de caso. Seu uso é justificado, principalmente, pela facilidade de representação e compreensão.

Quanto ao procedimento de verificação do trabalho, três estudos de caso são apresentados, evidenciando o consumo e a interligação dos recursos.

3 A publicação do índice Qualis

O primeiro passo em direção à publicação do índice Qualis foi o estabelecimento de uma linha temporal para o referido classificador, adotando-se o critério de replicação dos dados na representação de três triênios, como segue:

ISBN 978-85-61115-09-8

- **Qualis_2007**: representa os índices Qualis para os anos 2005 a 2007;
- **Qualis_2009**: representa os índices Qualis para os anos 2008 a 2010; e
- **Qualis_2013**: representa os índices Qualis para os anos 2011 a 2013.

Tabela 2 – Resumo do pré-processamento dos arquivos do índice Qualis

Material	Tuplas processadas	Tuplas validadas	% de perda
Qualis_2007	35.190	35.020	0,48
Qualis_2009	54.493	54.233	0,48
Qualis_2013	108.562	107.429	1,04

Fonte: elaborado pelos autores.

Ressalta-se que os arquivos enumerados passaram por um pré-processamento, convertendo-os em arquivo puramente texto e removendo tuplas em que não constavam, ou o ISSN de um elemento, ou uma classificação uma Qualis válida (diferentes de A1, A2, B1, B2, B3, B4, B5 e C). As tuplas remanescentes foram importadas para uma base de dados relacional. A Tabela 2 apresenta a quantidade de tuplas processadas, evidenciando a baixa perda de dados, ao considerar as restrições de consistência relatadas.

O processo de publicação do índice Qualis envolveu três estágios do ciclo de vida *LOD2 Stack*, sendo eles:

- **Extraction**: os dados foram recuperados do banco de dados relacional através de uma consulta, armazenando o resultado num arquivo formato CSV;
- **Evolution/Repair**: os dados do arquivo formato CSV foram convertidos para o formato RDF, melhorando sua expressividade ao adicionar elementos do metadados Dublin Core, através da ferramenta Google LODrefine; e
- **Storage/Querying**: os dados em formato RDF foram armazenados num *endpoint* baseado na ferramenta OpenLink Virtuoso RDF Store para permitir a execução dos estudos de caso.

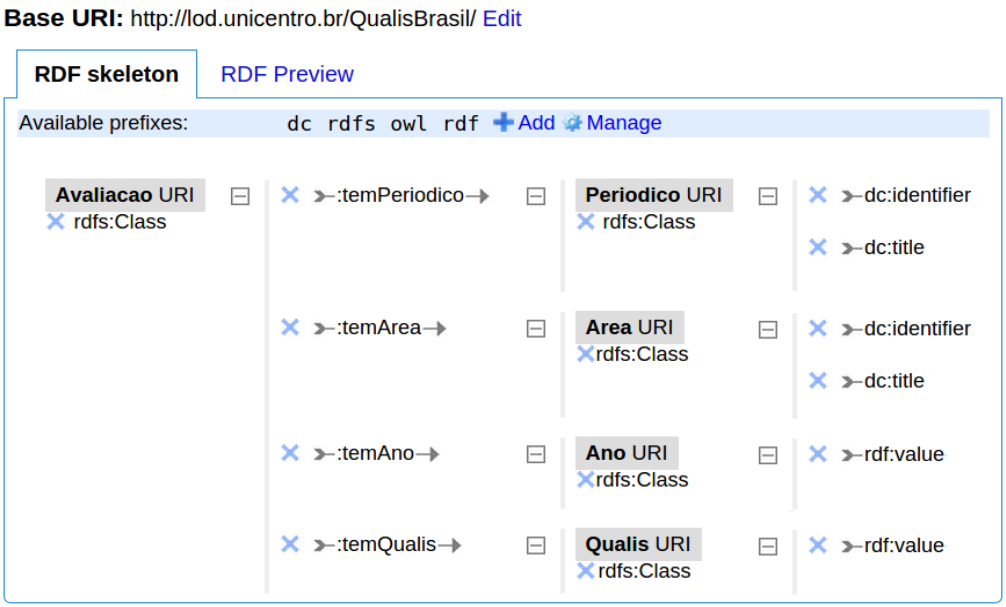
Para este processo pode-se fazer alguns apontamentos. Em soluções *Linked Open Data*, deve-se escolher vocabulários/modelos estabelecidos para facilitar a conversão de dados em recursos (URI - *Uniform Resource Identifier*) e posterior consumo (BAUER e KALTENBÖCK, 2014). Neste sentido, a organização dos dados publicados é baseada somente no uso de vocabulários e modelos amplamente difundidos na Web Semântica, o RDF Schema e o metadados Dublin Core. A Figura 3 apresenta a organização empregada.

Como exemplo dessa organização, a Listagem 1 codifica os recursos necessários para apresentar uma avaliação. Na referida listagem, destaca-se: (A) o prefixo "qb" para o uso dos recursos publicados a partir do endereço <http://lod.unicentro.br/QualisBrasil/>[17]; (B) a descrição do recurso para o periódico "*Expert Systems with Application*", atrelando seu ISSN, valendo-se do vocabulário Dublin Core; (C) a criação de um recurso de representação da área "Interdisciplinar", também de acordo o vocabulário Dublin Core; (D) a descrição de um recurso para vinculação do ano 2013; (E) a criação do recurso para representação do índice "Qualis A1"; e (F) a definição de um recurso que agrega a informação "o periódico *Expert Systems with Application* obteve o Qualis A1 na área Interdisciplinar no ano de 2013".

17 O endereço <http://lod.unicentro.br/QualisBrasil/> está em construção. Entretanto, os dados publicados estão armazenados em um grafo e acessíveis a partir do endpoint <http://space.sina.aksw.org/sparql> .

ISBN 978-85-61115-09-8

Figura 3 – Representação do vocabulário empregado



Fonte: elaborado pelos autores.

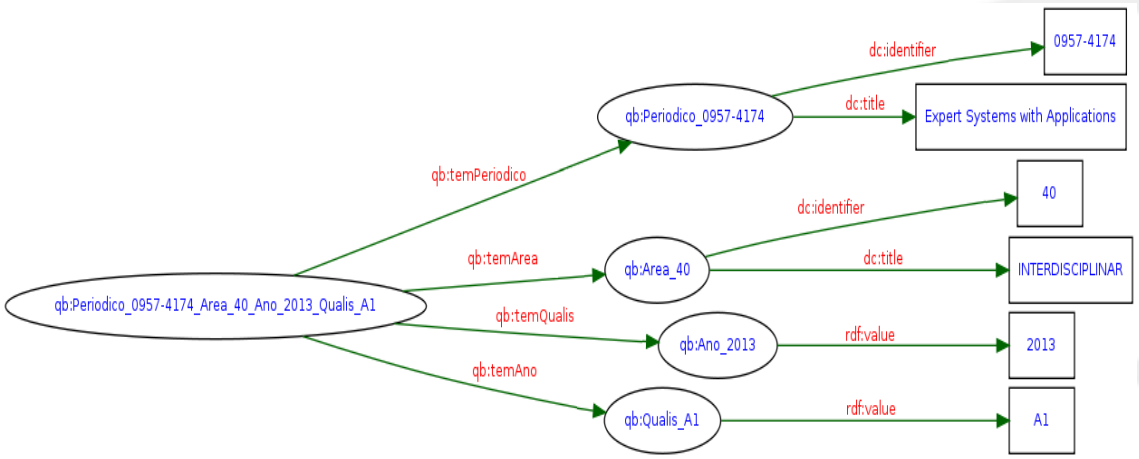
Listagem 1 – Código-fonte em RDF – exemplo de descrição de recursos

```
01 <?xml version="1.0" encoding="UTF-8"?>
02 <rdf:RDF
03   xmlns:dc="http://purl.org/dc/elements/1.1/"
04   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
05   xmlns:owl="http://www.w3.org/2002/07/owl#"
06   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
07 A xmlns:qb="http://lod.unicentro.br/QualisBrasil/">
08
09   <rdf:Description rdf:about="qb:Periodico_0957-4174">
10     <dc:identifier>0957-4174</dc:identifier>
11 B   <dc:title>Expert Systems with Applications</dc:title>
12   </rdf:Description>
13
14   <rdf:Description rdf:about="qb:Area_40">
15 C   <dc:identifier>40</dc:identifier>
16     <dc:title xml:lang="pt">INTERDISCIPLINAR</dc:title>
17   </rdf:Description>
18
19   <rdf:Description rdf:about="qb:Ano_2013">
20 D   <rdf:value rdf:datatype="http://www.w3.org/2001/XMLSchema#int">2013</rdf:value>
21   </rdf:Description>
22
23   <rdf:Description rdf:about="qb:Qualis_A1">
24 E   <rdf:value>A1</rdf:value>
25   </rdf:Description>
26
27   <rdf:Description rdf:about="qb:Periodico_0957-4174_Area_40_Ano_2013_Qualis_A1">
28     <qb:temPeriodico rdf:resource="qb:Periodico_0957-4174"/>
29 F   <qb:temArea rdf:resource="qb:Area_40"/>
30     <qb:temAno rdf:resource="qb:Ano_2013"/>
31     <qb:temQualis rdf:resource="qb:Qualis_A1"/>
32   </rdf:Description>
33 </rdf:RDF>
```

Fonte: elaborado pelos autores.

ISBN 978-85-61115-09-8

Figura 4 – Exemplo de grafo de representação para um recurso de avaliação



Fonte: elaborado pelos autores a partir de W3C (2014).

A Figura 4 representa a codificação dos recursos presentes na Listagem 1. Nesta figura são exemplificadas as regras de nomenclatura adotadas para cada recurso no momento da extração dos dados a partir do banco de dados relacional. Em suma, as regras aplicam a concatenação do nome do tipo do recurso a um código identificador, conforme codificado na Listagem 2.

Listagem 2 – Código-fonte em SQL – codificação parcial da recuperação de dados

```
01 select
02   concat('Periodico_',j.issnJournal,'_Area_',s.id,'_Qualis_',q.nameQualis) AS Avaliacao,
03   concat('Periodico_',j.issnJournal) AS Periodico,
04   concat('Area_',s.id) AS Area,
05   concat('Ano_',jsq.yearIndex) AS Ano,
06   concat('Qualis_',q.nameQualis) AS Qualis
07 from
08   journals j,
09   journalssubareasqualis jsq,
10   qualis q,
11   subareas s
12 where
13   jsq.fk_journals = jo.id and
14   jsq.fk_qualis   = q.id and
15   jsq.fk_subareas = s.id
```

Fonte: elaborado pelos autores.

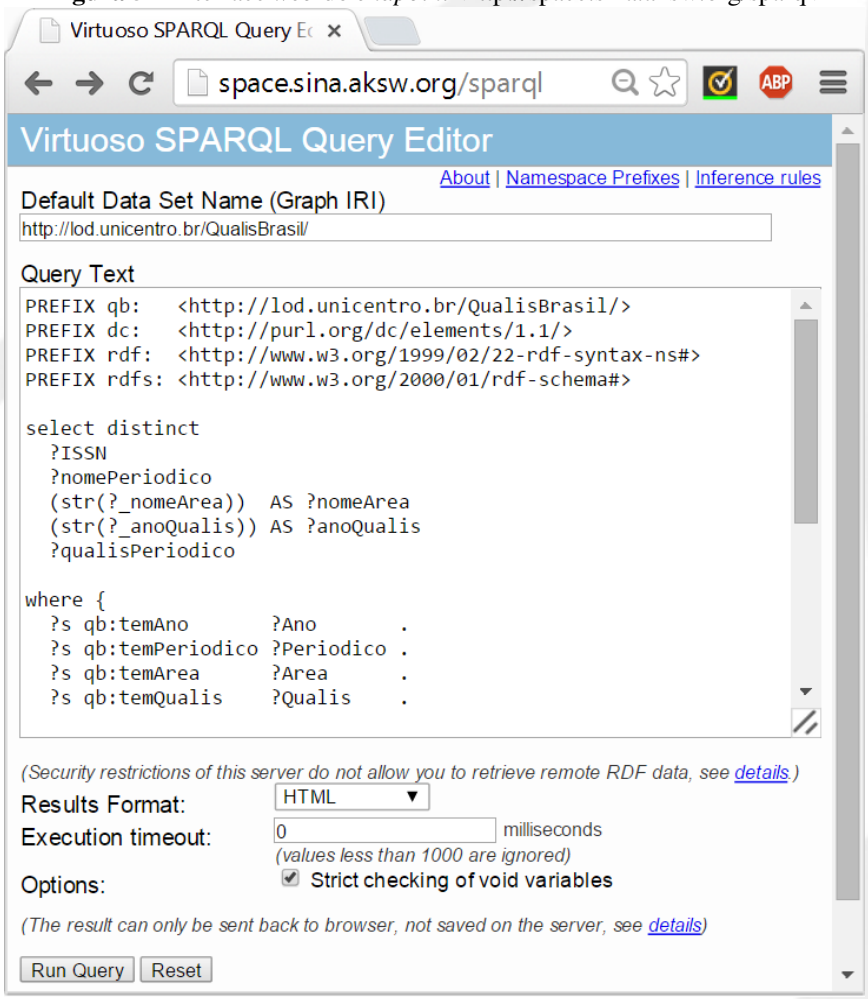
Uma vez apresentado o processo de publicação do índice Qualis, na continuação deste relato, são apresentados estudos de caso que exemplificam a utilização deste classificador de periódicos de forma aberta.

ISBN 978-85-61115-09-8

3 Estudos de Caso

Para verificar a publicação do índice Qualis, de acordo com os princípios do *Linked Open Data*, foram realizados três estudos de caso que exemplificam o consumo e a interligação de recursos em contextos cientométricos.

Figura 5 – Interface *web* do *endpoint*<http://space.sina.aksw.org/sparql>



Fonte: elaborado pelos autores.

Cabe ressaltar que o histórico do índice Qualis está disponível para consumo a partir do *endpoint*<http://space.sina.aksw.org/sparql>, conforme ilustrado na Figura 5.

3.1 Consumindo dados abertos: periódicos e seus qualis

Em *Linked Open Data*, dentre os princípios para consumir dados abertamente, prega-se a disponibilização dos dados de forma estruturada, utilizando um formato de representação não proprietário.

A Listagem 3, codificada na linguagem SPARQL, apresenta uma consulta para consumo do histórico do índice Qualis dos periódicos da área "Interdisciplinar", considerando o Qualis "A1" e o ano "2013". Na listagem, são destacados a padronização da nomenclatura para o uso dos dados abertos a partir do endereço <http://lod.unicentro.br/QualisBrasil/> e os

ISBN 978-85-61115-09-8

recursos a serem considerados na consulta. A Tabela 3 enumera exemplos de recursos recuperados.

Listagem 3 – Código-fonte em SPARQL – exemplo de consumo de dados do *endpoint*

```
PREFIX qb:   <http://lod.unicentro.br/QualisBrasil/>
PREFIX dc:   <http://purl.org/dc/elements/1.1/>
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select distinct
  ?ISSN
  ?nomePeriodico
  (str(?_nomeArea))  AS ?nomeArea
  (str(?_anoQualis)) AS ?anoQualis
  ?qualisPeriodico

where {
  ?s qb:temAno       ?Ano       .
  ?s qb:temPeriodico ?Periodico .
  ?s qb:temArea      ?Area      .
  ?s qb:temQualis    ?Qualis    .

  ?Ano    rdf:value ?_anoQualis      .
  ?Qualis rdf:value ?qualisPeriodico .

  ?Periodico dc:identifier ?ISSN     .
  ?Periodico dc:title ?nomePeriodico .
  ?Area      dc:title ?_nomeArea     .

  FILTER (
    ?Area   = qb:Area_40  &&
    ?Ano    = qb:Ano_2013 &&
    ?Qualis = qb:Qualis_A1
  )
} ORDER BY ?ISSN ?Ano LIMIT 5
```

Fonte: elaborado pelos autores.

Tabela 3 – Exemplos de recursos recuperados no processamento da consulta SPARQL – Listagem 3

ISSN	nomePeriodico	nomeArea	anoQualis	Qualis-Periodico
0001-0782	Communications of the ACM	INTERDISCIPLINAR	2013	A1
0001-2343	ARSP. Archiv für Rechts und Sozialphilosophie	INTERDISCIPLINAR	2013	A1
0001-3765	Anais da Academia Brasileira de Ciências (Impresso)	INTERDISCIPLINAR	2013	A1
0001-4966	The Journal of the Acoustical Society of America	INTERDISCIPLINAR	2013	A1
0001-6918	Acta Psychologica	INTERDISCIPLINAR	2013	A1

Fonte: elaborado pelos autores.

Cabe ressaltar que consultas similares de consumo de dados podem ser baseadas neste exemplo e integradas a diversas aplicações, ao utilizar ou ignorar os parâmetros de filtragem.

3.2 Interligando o índice Qualis à produção científica de um grupo de pesquisa

Uma das vantagens do *Linked Open Data* é a possibilidade em vincular dados de um sistema com os dados abertos de outras aplicações para que seja criado um contexto maior.

ISBN 978-85-61115-09-8

Como exemplo dessa potencialidade, a Listagem 4 codifica uma consulta SPARQL, reutilizando os dados disponíveis no *endpoint*<http://lod.unicentro.br/QualisBrasil/> para classificar parte da produção científica de um grupo de pessoas, destacando: (A) o prefixo "pp" para o contexto da produção em periódicos em outro grafo; (B) o prefixo "qb" para o contexto da publicação dos dados do índice Qualis; e (C) o recurso "Avaliacao" que interliga os dois contextos, possibilitando a reutilização dos dados do classificador.

Listagem 4 – Código-fonte em SPARQL – exemplo de criação de um contexto com os índices Qualis

```
01 PREFIX pp:   <http://lod.unicentro.br/ProducaoPeriodicos/>  (A)
02 PREFIX dc:   <http://purl.org/dc/elements/1.1/>
03 PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
04 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
05 PREFIX qb:   <http://lod.unicentro.br/QualisBrasil/>  (B)
06
07 SELECT DISTINCT
08   (str(?_labelGrupo) AS ?grupo)
09   (str(?_labelArea) AS ?area)
10   (str(?_labelAno) AS ?ano)
11   ?labelQualis
12   ?referencia
13
14 WHERE {
15   ?Grupo     dc:contributor  ?Professor    .
16   ?Professor pp:temArtigo    ?Artigo       .
17   ?Artigo    pp:temAvaliacao ?Avaliacao    .   (C)
18   ?Avaliacao qb:temAno       ?Ano          .
19   ?Avaliacao qb:temArea      ?Area         .
20   ?Avaliacao qb:temQualis    ?Qualis       .
21   ?Avaliacao qb:temPeriodico ?Periodico    .
22
23   ?Artigo    dc:identifier   ?referencia   .
24   ?Grupo     dc:title        ?_labelGrupo  .
25   ?Area      dc:title        ?_labelArea   .
26   ?Ano       rdf:value       ?_labelAno    .
27   ?Qualis    rdf:value       ?labelQualis  .
28
29   FILTER (
30     ?Grupo = pp:Grupo_3 &&
31     ?Area  = qb:Area_40 &&
32     ?Ano   = qb:Ano_2013
33   )
34 } ORDER BY DESC(?Ano) ?Qualis
```

Fonte: elaborado pelos autores.

Tabela 4 – Resultado do processamento da consulta SPARQL – Listagem 4

grupo	area	ano	labelQualis	referencia
Departamento de ...	INTERDISCIPLINAR	2013	A2	AGNER, L. T. W.; SOARES, I. W.; STADZISZ, P. C.; SIMAO, J. M. A Brazilian survey on UML and model-driven practices for embedded software development. Journal of Systems and Software. 2013.
Departamento de ...	INTERDISCIPLINAR	2013	A2	VENSKE, S. M. G. S.; GONÇALVES, R. A.; DELGADO, M. R. ADEMO/D: Multiobjective optimization by an adaptive differential evolution algorithm. Neurocomputing. 2013.
Departamento de ...	INTERDISCIPLINAR	2013	B1	DALL'AGNOL, J. M. H.; TACLA, C. A. A Method for Collaborative Argumentation in Merging Individual Ontologies. Journal of Universal Computer Science. 2013.
...	...	...	...	...

Fonte: elaborado pelos autores.

ISBN 978-85-61115-09-8

A Tabela 4 destaca os dados consumidos de <http://lod.unicentro.br/QualisBrasil/>, os quais foram interligados à produção científica de um grupo de pesquisadores presentes no grafo <http://lod.unicentro.br/ProducaoPeriodicos/>, possibilitando a classificação dos artigos publicados em periódicos da área Interdisciplinar.

O diferencial deste exemplo reside na liberdade de utilização de outros filtros. Por exemplo, é possível reclassificar a produção científica do mesmo grupo, de acordo as outras áreas do conhecimento, como "Ciência da Computação" ou "Ciências Sociais Aplicadas I".

3.3 Gerando informações cientométricas

De forma geral, obter informações cientométricas auxilia o reconhecimento de competências implícitas, a definição de ações para o desenvolvimento da ciência e de tecnologias, ou simplesmente, o entendimento institucional da evolução das pesquisas.

Listagem 5 – Código-fonte em SPARQL – exemplo de criação de consumo de informações cientométricas

```
01 PREFIX sap: <http://lod.unicentro.br/sumarizacaoArtigosPeriodicos/>
02 PREFIX qb:  <http://lod.unicentro.br/QualisBrasil/>
03 PREFIX pp:  <http://lod.unicentro.br/ProducaoPeriodicos/>
04 PREFIX dc:  <http://purl.org/dc/elements/1.1/>
05 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
06 PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
07 PREFIX scv: <http://purl.org/NET/scovo#>
08
09 select
10   (str(?_grupo) AS ?grupo)
11   (str(?_area)  AS ?area)
12   (str(?_ano)   AS ?ano)
13   ?qualis
14   (str(?_quantidade) As ?quantidade)
15
16 where {
17   ?Item   scv:dimension pp:Grupo_3 .
18   ?Item   scv:dimension qb:Area_40 .
19   ?Item   scv:dimension ?Ano       .
20   ?Item   scv:dimension ?Qualis    .
21
22   ?Item   rdf:value     ?_quantidade .
23   ?Ano    rdf:value     ?_ano        .
24   ?Qualis rdf:value     ?qualis      .
25
26   qb:Area_40 dc:title ?_area .
27   pp:Grupo_3 dc:title ?_grupo .
28
29   FILTER (
30     (?Ano != ?Qualis) &&
31     (datatype(?_ano) = xsd:int)
32   )
33 } ORDER BY ?Ano ?Qualis
```

Fonte: elaborado pelos autores.

Neste estudo de caso é exemplificado o uso de recursos disponibilizados no grafo <http://lod.unicentro.br/QualisBrasil/> no entendimento da evolução da produção científica.

O exemplo da Listagem 5 evidencia a utilização do vocabulário SCOVO (acrônimo "scv"). Nas linhas 17 a 20 são representadas as dimensões de um item de análise que remete à quantidade de publicações desenvolvida por um grupo de pessoas. O resultado desta consulta SPARQL pode ser parcialmente percebido na Tabela 5, destacando os dados consumidos a partir de <http://lod.unicentro.br/QualisBrasil/>.

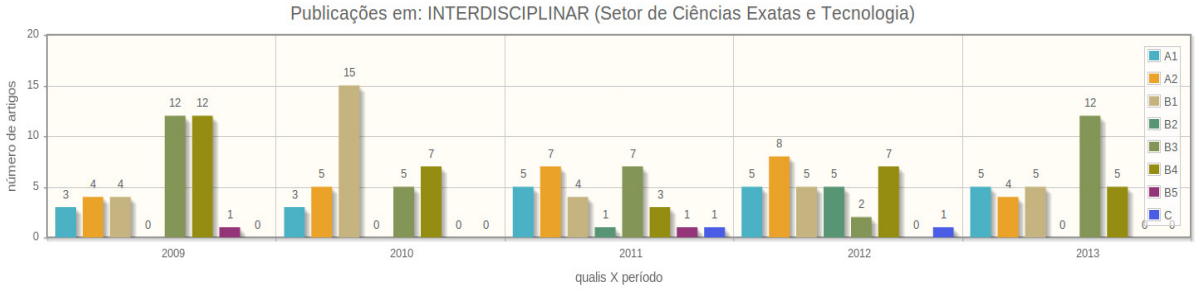
ISBN 978-85-61115-09-8

Tabela 5 – Resultado do processamento da consulta SPARQL – Listagem 4

grupo	area	ano	qualis	quantidade
Setor de Ciências ...	INTERDISCIPLINAR	2005	A2	3
Setor de Ciências ...	INTERDISCIPLINAR	2005	B1	2
Setor de Ciências ...	INTERDISCIPLINAR	2005	B2	1
Setor de Ciências ...	INTERDISCIPLINAR	2005	B3	1
Setor de Ciências ...	INTERDISCIPLINAR	2005	B4	7
Setor de Ciências ...	INTERDISCIPLINAR	2005	B5	1
...	...	...	...	...
Setor de Ciências ...	INTERDISCIPLINAR	2013	A1	5
Setor de Ciências ...	INTERDISCIPLINAR	2013	A2	4
Setor de Ciências ...	INTERDISCIPLINAR	2013	B1	5
Setor de Ciências ...	INTERDISCIPLINAR	2013	B3	12
Setor de Ciências ...	INTERDISCIPLINAR	2013	B4	5

Fonte: elaborado pelos autores.

Figura 6 – Exemplo de visualização de informações cientométricas



Fonte: elaborado pelos autores.

A título de ilustração, a Figura 6 apresenta os dados recuperados na forma de um histograma. Tal elemento de visualização pode ser um facilitador na interpretação de dados/informações cientométricos, servindo de apoio para justificar o esforço realizado no desenvolvimento de pesquisas e/ou tecnologias, por exemplo.

Considerações Finais

Neste relato de experiência é apresentado o primeiro resultado em direção do desenvolvimento de um "Modelo Tecnológico ao Compartilhamento de Dados para Estudos Cientométricos baseado em *Linked Open Data*".

Com o auxílio de estudos de caso foi possível demonstrar como a publicação de dados pode contribuir na coleta, na organização e no relacionamento de dados pertinentes, apoiando a realização de estudos bibliométricos ou cientométricos, observando que:

- o estudo de caso "*3.1 Consumindo dados abertos: periódicos e seus qualis*" exemplifica uma das formas para o consumo aberto de recursos em um ambiente *Linked Open Data*;
- o estudo de caso "*3.2 Interligando o índice Qualis à produção científica de um grupo de pesquisa*" demonstra como é possível consumir e interligar os recursos de <http://lod.unicentro.br/QualisBrasil/> em uma consulta; e
- o estudo de caso "*3.3 Gerando informações cientométricas*" também exemplifica o consumo e a vinculação dos recursos de <http://lod.unicentro.br/QualisBrasil/> em

ISBN 978-85-61115-09-8

um contexto, com o diferencial de utilizar vocabulário SCOVO para ampliar o teor informacional dos recursos perante a apresentação de um histograma.

Diante desses três estudos de caso, admite-se que a principal contribuição deste trabalho é a disponibilização de dados históricos do índice Qualis, de acordo com princípios do *Linked Open Data.* Disto, tais dados podem ser consumidos em diversos contextos e aplicações, principalmente, contribuindo para com pesquisas bibliométricas e cientométricas brasileiras.

Como trabalhos futuros pretende-se:

- expandir o modelo proposto, ao incorporar outros índices de classificação de periódicos, tais como os fatores de impacto *Journal Citation Reports* (JCR) e/ou *SCImago Journal Rank* (SJR);
- interligar os recursos do *endpoint*<http://lod.unicentro.br/QualisBrasil/> aos recursos disponibilizados no *endpoint* da DBpedia (<http://dbpedia.org>) para promoção de um contexto mais ampliado de busca;
- modelar os mecanismos para o consumo de recursos bibliométricos no escopo das universidades brasileiras, com o auxílio das ferramentas destacadas no ciclo de vida *LOD2 Stack*; e
- implementar uma interface amigável que facilite o consumo e a interligação dos recursos para com demais recursos em ambientes de *Linked Open Data*.

Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pela disponibilização de bolsa de estudos modalidade Estágio Pós-doutoral (Processo n° 18228/12-7) e ao grupo de pesquisa *Agile Knowlegde and Semantic Web* (Universidade de Leipzig) pela oportunidade da realização do estudo.

Referências

AKSW. **Agile Knowledge Engineering and Semantic Web**. Disponivel em: <http://aksw.org/About.html>. Acesso em: 29 de Agosto de 2013 10:00.

ARBOIT, A. E.; BUFREM, L. S.; GONZALEZ, J. A. M. A produção Brasileira em Ciência da Informação no exterior como reflexo de institucionalização científica. **Perspectivas em Ciência da Informação**, v. 16, n. 3, p. 75-92, 2011.

AUER, S. Introduction to LOD2. In: AUER, S.; BRYL, V.; TRAMP, S. (eds). **Linked Open Data – Creating Knowledge Out of Interlinked Data**. Springer, 2014.

AUER, S.; BÜHMANN, L.; DIRSCHL, C.; ERLING, O.; HAUSENBLAS, M.; ISELE, R.; LEHMANN, J.; MARTIN, M.; MENDES, P. N.; van NUFFELEN, B.; STADLER, C.; TRAMP, S.; WILLIAMS, H. Managing the Life-Cycle of Linked Data with the LOD2 Stack. **Lecture Notes in Computer Science**, v. 7650, p 1-16, 2012.

AUER, S., LEHMANN, J., NGOMO, A. N. N., ZAVERI, A. Introduction to Linked Data and Its Lifecycle on the Web. **Lecture Notes in Computer Science**, v. 8067, p. 1-90, 2013.

BAUER, F.; KALTENBÖCK, M. **Linked Open Data: The Essentials**. Disponível em: <http://www.semantic-web.at/LOD-TheEssentials.pdf>. Acesso em: 19 de Outubro de 2014 19:00.

ISBN 978-85-61115-09-8

CUNHA, M. B. da; CAVALCANTI, C. R. de O. **Dicionário de biblioteconomia e arquivologia**. Brasília: Briquet de Lemos, 2008.

DCMI. **DCMI Metadata Terms**. Disponível em: <http://dublincore.org/documents/dcmi-terms/>. Acesso em: 27 de Julho de 2014 14:00.

KONUR, O. The Evaluation of the Global Research on the Education: A Scientometric Approach. **Procedia - Social and Behavioral Sciences**, v. 47, p. 1363-1367, 2012.

LATTES. **Plataforma Lattes**. Disponível em: <lattes.cnpq.br>. Acesso em: 25 de Agosto de 2013 10:00.

LINKED DATAa. **Linked Data – Connect Distributed Data across the Web**. Disponível em: <http://linkeddata.org>. Acesso em: 28 de Agosto de 2012 16:00.

LINKED DATAb. **Linked Data – Design Issues**. Disponível em: <http://www.w3.org./DesignIssues/LinkedData.html>. Acesso em: 28 de Agosto de 2013 17:00.

LOD2. **Linked Data Stack | LOD2 Technology Stack**. Disponível em: < http://stack.lod2.eu/blog/>. Acesso em: 20 de Maio de 2014 14:00.

LUCHS, A. Profile of Brazilian scientific production on A/H1N1 pandemic influenza. **Ciênc. saúde coletiva**, v. 17, n. 6, p. 1629-1634, 2012.

MENA-CHALCO, J. P.; CESAR-JR, R. M. scriptLattes: An open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, vol. 15, n. 4, p. 31-39, 2009.

HAUSENBLAS, M.; HALB, W.; RAIMOND, Y.; FEIGENBAUM, L.; AYERS, D. SCOVO: Using Statistics on the Web of Data. In.: Semantic Web in Use Track of the 6th European Semantic Web Conference, Heraklion, Grécia, 2009. ESWC2009 Proceedings..., Heraklion, 2009.

MOURA, A. M. M. de; CAREGNATO, S. E. Co-autoria em artigos e patentes: um estudo da interação entre a produção científica e tecnológica. **Perspectivas em Ciência da Informação**. v. 16, n. 2, p. 153-167, 2011.

MYSQL. **MySQL :: MySQL Downloads**. Disponível em: <http://dev.mysql.com/downloads/>. Acesso em: 20 de Maio de 2014.

SANTOS, R. N. M. dos; KOBASHI, N. Y.. Bibliometria, cientometria, infometria: conceitos e aplicações. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v.2, n.1, p. 155-172. 2009.

W3C. **W3C RDF Validation Service**. Disponível em: <http://www.w3.org/RDF/Validator/>. Acesso: 24 de Julho de 2014.

WEBQUALIS. **Sistema WebQualis - Portal Capes**. Disponível em: <http://qualis.capes.gov.br/webqualis/principal.seam>. Acesso em: 25 de Agosto de 2013 10:00.

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

Desenvolvimento de Web APIs RESTful Semânticas

Ivan Salvadori
ivan.salvadori@posgrad.ufsc.br

Frank Siqueira
frank@inf.ufsc.br

Resumo
Os princípios arquiteturais REST estão sendo amplamente adotados nas implementações de sistemas distribuídos disponíveis na Web, cujo principal objetivo é o intercâmbio de dados entre aplicações. No entanto, devido à falta de padrões e diretrizes, cada implementação segue uma linha de desenvolvimento diferente. Outro desafio é a falta de suporte ao uso de controles hipermídia em representações JSON. Este trabalho propõe um *framework* para o desenvolvimento de Web APIs RESTful que suportam controles hipermídia no formato JSON. O *framework* realiza anotações semânticas nos recursos Web através de associações de propriedades e operações com termos de vocabulários controlados, possibilitando menor acoplamento entre Web APIs e aplicações cliente.

Palavras-chave: Web API RESTful. Descrição Semântica. Dados Ligados.

Abstract
The REST architectural principles are being widely adopted to implement Web-based distributed systems, aimed at allowing data interchange among applications. Nonetheless, due to the lack of standards and guidelines, different Web APIs follow the most diverse design approaches. Another challenging issue in this area is the lack of support for hypermedia controls in data represented in the JSON format. This work proposes a framework for RESTful Web API development with support for hypermedia controls in JSON data. The proposed framework adds semantic annotations to Web resources by associating properties and operations with concepts defined in controlled vocabularies, fostering loose coupling between client applications and Web APIs.

Key Words: RESTful Web API. Semantic Description. Linked Data.

Introdução

Com a necessidade de integração de dados cada vez maior, é fundamental o uso de mecanismos eficientes para troca de informações entre sistemas. Utilizar Web APIs possibilita que diferentes aplicações troquem informações com um nível mais baixo de acoplamento. Com o uso de Web APIs é possível compartilhar informações em escala global através da Web. Independentemente da escala, a integração de dados é melhor realizada quando as informações estão descritas semanticamente, fato que transforma a Web de recursos isolados na Web de Dados (BERNERS-LEE; HENDLER; LASSILA, 2001).

As informações manipuladas por Web APIs possuem formatos mais adequados para serem consumidos por aplicações, uma vez que os seres humanos não são os principais consumidores. Os formatos de dados mais utilizados são XML e JSON (RICHARDSON; AMUNDSEN; RUBY, 2013). Para oferecer suporte a *Linked Data*, surge a recomendação W3C denominada JSON-LD (LANTHALER; GüTL, 2012). Embora JSON-LD represente um grande avanço, limita-se ao desenvolvimento de Web APIs somente-leitura, pois não tem

ISBN 978-85-61115-09-8

suporte para descrever a semântica necessária para alteração de recursos Web (RICHARDSON; AMUNDSEN; RUBY, 2013). Surge então Hydra (LANTHALER, 2013) como uma proposta para desenvolver Web APIs RESTfulcom suporte a controles hipermídia, permitindo a leitura e alteração de recursos. Controles hipermídia, entre os quais pode-se destacar os *links* e formulários HTML, permitem a navegação e transformação de estados de recursos.

Os *frameworks* disponíveis atualmente para desenvolvimento de Web APIs baseadas nos princípios arquiteturais REST possuem suporte limitado para adição de informação semântica e de controles hipermídia. RestML,proposto por SANCHEZ, OLIVEIRA e FORTES (2014), é voltado à modelagem de *Web Services* REST através de um *profile* UML e à geração automática do código fonte da aplicação, sem qualquer suporte a descrição semântica ou a controles hipermídia. Spring-HATEOAS (SPRING, 2013) e Apache Isis (HAYWOOD, 2013) suportam controles hipermídia, mas não oferecem suporte para descrição semântica de dados. Por outro lado, JOHN e RAJASREE (2012) abordam a descrição semântica da documentação de Web APIs.

Este trabalho fornece suporte ferramental para o desenvolvimento de Web APIs RESTful semânticas que utilizam o formato de dados JSON. O suporte ferramental é constituído por um *framework* que combina suporte a controles hipermídia com Web Semântica e *Linked Data*. O *framework* proposto realiza a geração automática da documentação em dois formatos. O primeiro formato utiliza a estrutura de documentação Hydra, adequado para ser interpretado por outras aplicações. Outro formato de documentação disponibilizado é o HTML, adequado para guiar os desenvolvedores de aplicações cliente.

O restante do artigo está organizado da seguinte forma: a Seção 1 apresenta os fundamentos necessários para o entendimento do artigo. O *framework* proposto é descrito na Seção 2. A Seção 3 apresenta os trabalhos relacionados. A última seção apresenta as conclusões dos autores e descreve as possibilidades de trabalhos futuros.

1 Fundamentos Tecnológicos

Este trabalho é baseado em Web APIs REST com suporte a descrição semântica, *Linked Data* e controles hipermídia proporcionados através de JSON-LD e Hydra, além da especificação JAX-RS. Estas tecnologias serão descritas ao longo desta seção.

1.1 REST

REST (*Representational State Transfer*) é uma coleção de princípios e restrições arquiteturais para o desenvolvimento de aplicações distribuídas na Web. Adota o paradigma cliente-servidor, onde as requisições partem inicialmente do cliente e são respondidas pelo servidor. O servidor não deve guardar informações sobre requisições de seus clientes, pois assume um comportamento *stateless* (FIELDING, 2000).

Recursos formam a base dos princípios REST e podem ser qualquer informação que se deseje tornar acessível a clientes remotos, e que são endereçados através de um identificador único (URI). Recursos podem ser uma lista de filmes em cartaz em um cinema, comentários de um blog, uma página pessoal ou um perfil de um usuário de uma rede social, por exemplo. Um recurso pode ser identificado por diversas URIs, mas uma URI endereça apenas um recurso. Recursos podem ser representados de diferentes maneiras, ou seja, podem ter diferentes representações. A representação é uma amostra do estado do recurso em um determinado momento do tempo. O recurso jamais é acessado diretamente, mas através de uma representação. Sendo assim, umaURI está sempre associada a pelo menos uma

ISBN 978-85-61115-09-8

representação. As representações podem assumir vários formatos, por exemplo, HTML, XML, JSON, etc.

O estilo arquitetural REST adota o princípio HATEOAS (*Hypermedia as the Engine of Application State*). Define que as mudanças de estado da aplicação devem ser guiadas através de controles hipermídia (FIELDING, 2000). Entende-se por aplicação o conjunto de recursos e seus respectivos estados manipulados por um cliente. Controles hipermídia podem assumir a forma de *links*, que guiam a navegação entre diferentes recursos ou formulários, oferecem mecanismos para criação ou alteração de informações. HATEOAS implica que as representações de recursos, além de conter dados, devem conter controles hipermídia com transições válidas para um determinado estado.

1.2 JAX-RS

A Especificação JAX-RS (*Java API for RESTful Services*) define um conjunto de APIs para o desenvolvimento de serviços Web, empregando a linguagem de programação Java, que obedecem aos princípios arquiteturais REST (ORACLE, 2013). Proporciona um conjunto de anotações capazes de expor classes POJOs (*Plain Old Java Objects*) como recursos, suporte ao protocolo HTTP e independência de formatos (diferentes representações).

A Figura 1 ilustra um exemplo de anotações JAX-RS. Através da anotação *@Path* (linha 1) aplicada sobre uma classe, cria-se um recurso Web, dados de uma pessoa por exemplo, que pode ser acessado através da URI */pessoa/{id}*. Para tornar o endereçamento dos recursos mais flexível, umaURI pode conter variáveis que são mapeadas para propriedades. A variável é mapeada para uma propriedade da classe através da anotação *@PathParam,* que especifica o nome da variável(linha 4).

Quatro métodos são implementados para permitir a manipulação dos recursos, que são anotados com *@GET*, *@POST*, *@PUT* e *@DELETE* (linhas 15, 21, 27 e 34) e criam a correspondência entre os tipos de requisições HTTP e a execução dos métodos anotados. As anotações *@Consumes* e *@Produces* são responsáveis por mapear os tipos de dados de entrada e saída, respectivamente, aceitos por cada método. Os argumentos que constituem a assinatura dos métodos (linhas 23 e 30) são automaticamente mapeados do formato aceito descritos pela anotação *@Consumes* para o argumento referenciado. Cada método retorna um objeto do tipo *Response*, que representa a resposta a ser enviada pelo protocolo HTTP. O objeto *Response* retornado pelo método contém o código de status HTTP para a requisição e, opcionalmente, uma representação de recurso vinculado, denominado *entity*. Essa representação, quando presente, é automaticamente mapeada para o formato indicado pela anotação *@Produces*. Caso mais de um formato seja suportado, é escolhido o formato de preferência do cliente, indicado no cabeçalho da requisição HTTP.

ISBN 978-85-61115-09-8

Figura 1 – Classe anotada com JAX-RS

```
1   @Path("/pessoa/{id}")
2   publicclass PessoaResource {
3
4       @PathParam("id")
5   private Long id;
6
7       private String nome;
8
9   private String sobrenome;
10
11  private String imagem;
12
13  private List<PessoaResource> amigos;
14
15      @GET
16      @Produces(MediaType.APPLICATION_JSON)
17  public Response carregar() {
18          //codigo para carregar uma pessoa
19      }
20
21      @POST
22      @Consumes(MediaType.APPLICATION_JSON)
23  public Response criar(Pessoa p) {
24          //codigo para criar pessoa p
25      }
26
27      @PUT
28      @Produces(MediaType.APPLICATION_JSON)
29      @Consumes(MediaType.APPLICATION_JSON)
30  public Response atualizar(Pessoa p) {
31          //codigo para alterar pessoa p
32      }
33
34      @DELETE
35  public Response remover() {
36          //codigo para remover uma pessoa
37      }
38  }
```

Fonte: do autor.

1.3 JSON-LD

JSON-LD é uma maneira leve para descrever *Linked Data* no formato JSON. Adiciona informações semânticas em documentos JSON existentes, sem necessidade de grandes esforços. Foi projetado para proporcionar simplicidade, compatibilidade e expressividade. JSON-LD é compatível com JSON, ou seja, todo documento JSON-LD é um documento JSON válido. Essa compatibilidade permite que as bibliotecas e analisadores atuais sejam reutilizados. Seus principais objetivos são possibilitar a construção de *Web Services* com recursos semânticos, além de armazenamento de *Linked Data* em *engines* baseadas em JSON (LANTHALER; GüTL, 2012).

JSON-LD permite criar contextos compostos por classes e propriedades de vocabulários controlados. Vocabulários controlados organizam e agrupam termos que podem

ISBN 978-85-61115-09-8

ser associados a um domínio específico (SVENONIUS, 1986). Formam uma coleção de termos organizados que facilitam o acesso à informação, podendo ser utilizados para descrever informações a fim de facilitar sua recuperação (WARNER, 2002) e interpretação. Os vocabulários são utilizados como fonte de metadados que descrevem o conteúdo dos recursos. Exemplos de vocabulários controlados são *schema.org,FOAF e dbpedia*, que fornecem significados para conceitos organizados em classes, propriedades e relacionamentos.

JSON-LD permite associar propriedades com termos de diferentes fontes semânticas, além da possibilidade de associar o documento a um conjunto de classes que vincula um significado para o documento como um todo. Para associar vocabulários ao documento, JSON-LD disponibiliza a notação *@context.* JSON-LD permite representar valores de propriedades como *links*, diferenciando *URLs* de simples valores textuais. Para definir que uma determinada propriedade corresponde a um *link* usa-se a notação *"@type": "@id"*. A informação descrita semanticamente em JSON-LD possui *URLs* para que o consumidor possa obter mais informações sobre determinado significado. Relacionamentos também podem possuir semântica associada, de modo a estabelecer uma relação semântica entre recursos. Enriquecer semanticamente informações com termos compartilhados e conhecidos resulta em integrações mais ricas, flexíveis e menos vulneráveis a modificações.

A Figura 2 ilustra um exemplo de documento representado com JSON-LD. Entre as linhas 1 a 13 é definido o contexto do documento. O documento é associado ao conceito de *Person* presente nos vocabulários*schema* e *foaf,* através da anotação *@type* (linha 3). Os vocabulários utilizados são declarados nas linhas 4 e 5, e possuem um prefixo e o endereço do vocabulário. As propriedades são associadas aos termos dos vocabulários declarados (linhas 6 a 12). Essa associação permite obter o significado da propriedade através da consulta ao vocabulário corresponde. Para obter mais informações sobre a propriedade *schema:name*, por exemplo, basta visitar a URL do vocabulário juntamente com o termo desejado (*http://schema.org/name*). A anotação *@type* com valor *@id* (linha 10) identifica que a propriedade é representada por um *link*, e não deve ser interpretada como valor textual. A propriedade *amigos* (linha 12) possui o significado *foaf:knows* que descreve um determinado relacionamento entre recursos. A propriedade *@id* (linha 14) apresenta o identificador único do documento, geralmente representado pela URL do próprio recurso. Com o contexto definido, cada propriedade do documento está vinculada a termos de vocabulários controlados.

A combinação entre o estilo arquitetural REST e os princípios de *Linked Data* proporciona grande avanço para a troca de informações entre Web APIs e agentes autônomos (clientes). Essa combinação baseia-se no enriquecimento semântico de dados, de forma a tornar a informação compreensível aos agentes. Nesse cenário surge o Hydra (LANTHALER, 2013), que provê um vocabulário capaz de representar o significado de transações de estados dos recursos. Isso significa que clientes de Web APIs passam a ter informações suficientes para realizar requisições que modificam o estado dos recursos, através da extensão do formato JSON-LD.

ISBN 978-85-61115-09-8

Figura 2 – Exemplo de documento JSON-LD

```
1   {
2     "@context": {
3       "@type": ["schema:Person", "foaf:Person"],
4       "schema": "http://schema.org/",
5       "foaf": "http://xmlns.com/foaf/0.1/",
6       "nome": "schema:givenName",
7       "sobrenome": "schema:familyName",
8       "foto": {
9         "@id": "schema:image",
10        "@type": "@id"
11      },
12      "amigos": "foaf:knows"
13    },
14    "@id": "http://web-api.com/user/1",
15    "nome": "Lucky",
16    "sobrenome": "Luke",
17    "foto": "http://web-api.com/lucky.jpg",
18    "amigos": [
19      "http://web-api.com/user/2",
20      "http://web-api.com/user/3"
21    ]
22  }
```

Fonte: do autor.

1.4 Hydra

Embora JSON-LD represente um avanço para a realização de controles hipermídia, apenas *hiperlinks* podem ser vinculados a representações. As demais operações que modificam o estado do recurso não podem ser representadas (RICHARDSON; AMUNDSEN; RUBY, 2013). Hydra propõe uma alternativa que possibilita o uso mais amplo de controles hipermídia, através da ampliação dos conceitos do JSON-LD e de um vocabulário específico para tratar requisições. Adota o conceito de classes para descrever recursos, que contém as propriedades e as operações suportadas. O suporte oferecido pelo Hydra permite a troca de informações no formato JSON, utilizando o contexto semântico do JSON-LD com a adição das instruções necessárias para execução de operações sobre o recurso Web. Além de enriquecer as informações disponibilizadas pela Web API, Hydra oferece uma maneira de descrever todos os recursos gerenciados pelo servidor, através da documentação da Web API.

A Figura 3 ilustra um exemplo de documento representado no formato Hydra. O contexto do documento é definido de forma idêntica ao JSON-LD, porém novas notações estendem o poder de representação de controles hipermídia. Através da notação *operations*, é possível associar operações sobre o recurso a métodos HTTP. Três operações são definidas e permitem a criação, alteração e a remoção do recurso (linhas 8 a 21). Hydra permite diferenciar as naturezas de operações através de palavras reservadas. As operações de criação, alteração e remoção são denominadas Create*ResourceOperation, ReplaceResourceOperation* e *DeleteResourceOperation* (linhas 10, 14 e 18) respectivamente. Hydra permite descrever todos os requisitos necessários para a execução de uma operação. Cada operação informa também o método HTTP que deve ser utilizado para executá-la (linhas 11, 15 e 19). A URL base para execução é informada pela propriedade *@id* do documento (linha 3). Hydra suporta operações que alteram o estado dos recursos Web, dessa forma, amplia a expressividade do JSON-LD e torna-se uma alternativa mais completa para o desenvolvimento de Web APIs.

ISBN 978-85-61115-09-8

Figura 3 – Exemplo de documento com controles hipermídia Hydra

```
1   {
2     "@context": {...},
3     "@id": "http://web-api.com/user/1",
4     "nome": "Lucky",
5     "sobrenome": "Luke",
6     "foto": "http://web-api.com/lucky.jpg",
7     "amigos": [...],
8     "operations": [
9   {
10        "@type": "CreateResourceOperation",
11        "method": "POST"
12      },
13      {
14        "@type": "ReplaceResourceOperation",
15        "method": "PUT"
16      },
17      {
18        "@type": "DeleteResourceOperation",
19        "method": "DELETE"
20      }
21    ]
22  }
```

Fonte:do autor.

2 Framework JAX-SRS

Frameworks que ofereçam suporte adequado para a implementação de serviços Web são comumente utilizados para garantir a produtividade da implementação e a qualidade do *software*. Representações com suporte a controles hipermídia (LANTHALER; GüTL, 2012), relacionamentos entre recursos através de *links*, suporte a coleções e descrição semântica (BERNERS-LEE; HENDLER; LASSILA, 2001) são alguns passos rumo a implementações de Web APIs verdadeiramente RESTful. O *framework* proposto, denominado JAX-SRS (*Java Framework for Semantic RESTful Services*), incorpora uma série de conceitos de modelagem de *software*, que resultam em serviços Web que seguem os princípios REST, adicionados de descrição semântica e de controles hipermídia, objetivando uma implementação flexível, facilmente expansível e escalável.

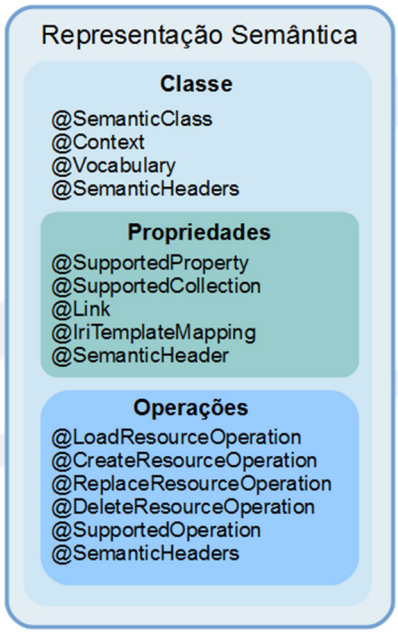
O *framework* proposto tem como principal objetivo possibilitar o desenvolvimento de Web APIs RESTful com suporte a controles hipermídia e *Linked Data* para representações baseadas em JSON. Com base na especificação JAX-RS, oferece suporte ao desenvolvimento através de duas vertentes. A primeira é a adição de controles hipermídia nas representações disponibilizadas aos clientes da Web API. A segunda é a construção da documentação da Web API, que descreve as representações dos recursos, suas propriedades e operações suportadas, enriquecidas com significado semântico. JAX-SRS disponibiliza um conjunto de anotações que permite a descrição semântica das representações com suas propriedades e operações.

2.1 Conjunto de Anotações

O *framework* JAX-SRS possui um conjunto de anotações que adiciona significado semântico sobre o elemento anotado. O conjunto de anotações é complementar às anotações JAX-RS e mantém a compatibilidade com a especificação. As anotações podem ser aplicadas sobre classes, atributos ou métodos (Figura 4).

ISBN 978-85-61115-09-8

Figura 4 – Conjunto de anotações JAX-SRS



Fonte:do autor.

Dentre as anotações que podem ser aplicadas sobre classes estão: *@SemanticClass*, que associa a classe como uma representação REST; *@Context*, que permite a construção do contexto semântico da representação; *@Vocabulary*, que associa um ou diversos vocabulários ao contexto; e *@SemanticHeaders*, que permite definir cabeçalhos HTTP necessários para a execução das operações declaradas.

As anotações *@SupportedProperty* e *@Link* são aplicadas sobre os atributos da classe e vinculam as propriedades da representação com termos relacionados de vocabulários declarados no contexto semântico. A diferença entre as duas anotações é a forma como a propriedade é apresentada; enquanto *@SupportedProperty* sinaliza que o valor da propriedade é parte integrante da representação, *@Link* indica que o valor da propriedade é um *hiperlink*. A anotação *@SupportedCollection* é responsável por definir que a propriedade associada corresponde a uma lista de recursos. Através da anotação *@IriTemplateMapping* é possível adicionar significado semântico a variáveis presentes em URIs. Da mesma forma que propriedades podem mapear variáveis de URI, propriedades podem também ser associadas a cabeçalhos HTTP. A anotação *@SemanticHeader* permite definir um significado semântico para o mapeamento de cabeçalhos HTTP.

Dentre as anotações que podem ser aplicadas sobre os métodos de classes estão: *@LoadResourceOperation*, *@CreateResourceOperation*, *@ReplaceResourceOperation* e *@DeleteResourceOperation*. Essas anotações, além de possibilitar a vinculação de significado semântico, definem a natureza das operações, que pode ser de carregamento, criação, atualização ou remoção do recurso, respectivamente. A anotação *@SupportedOperation* apenas vincula significado semântico à operação, sem adicionar nenhum significado relativo à sua natureza. Por fim, a anotação *@SemanticHeaders* permite definir os cabeçalhos HTTP necessários para a execução da operação anotada.

A Figura 5 apresenta um exemplo de aplicação de anotações do *framework* JAX-SRS. Na linha 2, o recurso é associado às classes *schema:Person* e *foaf:Person*. O contexto semântico inicia com a anotação *@Context* juntamente com dois vocabulários, que são constituídos por um prefixo *"value"* e uma URL para o vocabulário (linhas 3 e 4). O recurso possui uma variável em sua URL, que é mapeada para a propriedade *id* (linha 9). O valor

ISBN 978-85-61115-09-8

semântico dessa variável é descrito através da anotação *@IriTemplateMapping* associado a *schema:taxID*. As propriedades *nome* e *sobrenome* são anotadas com *@SupportedProperty* e são associadas aos termos *schema:name* e *schema:givenName*, respectivamente (linhas 11 a 15). A propriedade *imagem* (linha 18) deve ser apresentada em forma de *link,* enquanto a propriedade *amigos* (linha 21) corresponde a uma coleção de recursos.

Figura 5 – Exemplo de anotações JAX-SRS

```
1  @Path("/pessoa/{id}")
2  @SemanticClass(id = {"schema:Person", "foaf:Person"})
3  @Context({@Vocabulary(value = "schema", url = "http://schema.org/"),
4            @Vocabulary(value = "foaf", url = "http://xmlns.com/foaf/0.1/") })
5  publicclass PessoaResource {
6
7      @PathParam("id")
8  @IriTemplateMapping(id = "schema:taxID")
9  private String id;
10
11     @SupportedProperty(id = "schema:name")
12 private String nome;
13
14     @SupportedProperty(id = "schema:givenName")
15 private String sobrenome;
16
17     @Link(id = "schema:image")
18 private String imagem;
19
20     @SupportedCollection(id = "foaf:knows")
21 private List<PessoaResource>amigos;
22
23     @GET
24     @Produces("application/ld+json")
25     @LoadResourceOperation(returnedClass = PessoaResource.class)
26 public Response load() {...}
27
28     @POST
29     @Consumes("application/ld+json")
30     @CreateResourceOperation(expectedClass = PessoaResource.class)
31     @SemanticHeaders(@SemanticHeader(id = "HTTP BASIC", name = "Authorization"))
32 public Response create(PessoaResource p) {...}
33
34     @PUT
35     @Consumes("application/ld+json")
36     @ReplaceResourceOperation(expectedClass = PessoaResource.class)
37     @SemanticHeaders(@SemanticHeader(id = "HTTP BASIC", name = "Authorization"))
38 public Response update(PessoaResource p) {...}
```

Fonte:do autor.

Além das propriedades, um recurso pode possuir operações que carregam ou alteram os valores de suas propriedades. O exemplo da Figura 5 declara quatro métodos, que representam operações sobre o recurso (linhas 26, 32, 38 e 43). Além das anotações JAX-RS *@GET, @POST, @PUT* e *@DELETE*, e seus respectivos *@Consumes* e *@Produces*, os métodos possuem anotações do *framework* JAX-SRS. O método *load* (linha 26) é anotado com *@LoadResourceOperation* com o parâmetro *returnedClass = PessoaResource.class*. Isso implica que a resposta da requisição HTTP GET contém uma representação do recurso *PessoaResource* no formato *application/ld+json*, correspondente ao formato JSON-LD. O método *create* (linhas 32) é anotado com *@CreateResourceOperation* com o parâmetro

ISBN 978-85-61115-09-8

expectedClass = PessoaResource.class. Esta operação é responsável pela criação de um novo recurso do tipo *PessoaResource* com os valores informados na representação, no formato JSON-LD, presente na requisição HTTP POST. O método *update* (linha 38) é semelhante ao método *create*, porém o resultado é a alteração dos valores atuais pelos valores da representação informados na requisição HTTP PUT. O método *delete* (linha 43) é anotado com *@DeleteResourceOperation* e deve ser interpretado como responsável pela remoção do recurso correspondente a uma determinada URL. Os métodos *create, updade* e *delete* também recebem a anotação *@SemanticHeaders*, que implica que a execução da operação considera a presença do cabeçalho HTTP *Authorization*. É possível implementar operações que não possuem as características de carregamento, criação, atualização ou remoção. Essas operações realizam atividades mais complexas, e são suportadas pelo *framework* através da anotação *@SupportedOperation,* podendo ser associadas a um significado que descreve semanticamente a operação.

2.2 Módulos do Framework

O *framework* JAX-SRS é constituído pelo módulo de suporte a controles hipermídia e também pelo módulo de geração de documentação da Web API. Esse módulo adiciona controles hipermídia às representações disponibilizadas pelo *framework*. Os controles hipermídia correspondem às operações do recurso, e são adicionados diretamente na representação, juntamente com as propriedades. Essa característica é importante pois proporciona ao consumidor da informação conhecer as operações permitidas para um determinado recurso, fato que amplia o seu poder de decisão. Além dos controles hipermídia, é adicionado o contexto, que apresenta a estrutura e a descrição semântica do recurso. O módulo de geração de documentação é responsável por disponibilizar a estrutura dos recursos, denominada *supportedClasses*, bem como a descrição semântica das propriedades (Figura 6) e operações (Figura 7) em um único documento no formato Hydra.

A documentação da Web API no formato Hydra é adequada para ser interpretada por outras aplicações. Embora seja legível, não é o formato mais adequado para seres humanos. O módulo de geração da documentação gera uma documentação da Web API em formato HTML, ideal para guiar os desenvolvedores de aplicações cliente. A Figura 8 mostra um exemplo da documentação HTML gerada pelo *framework*, que apresenta todos os recursos semânticos identificados (*Supported Classes*), além de detalhes como contexto semântico, propriedades e operações suportadas. A documentação gerada apresenta ao desenvolvedor todos os detalhes estruturais e semânticos do recurso, além das informações necessárias para a execução das operações.

ISBN 978-85-61115-09-8

Figura 6 – Exemplo de Documentação Hydra para Web APIs

```
{
  "supportedClasses": [
    {
      "@context": {
        "schema": "http://schema.org/",
        "foaf": "http://xmlns.com/foaf/0.1/"
      },
      "@id": ["schema:Person", "foaf:Person"],
      "supportedProperties": [
        {
          "@id": ["schema:name"],
          "property": "nome", "@type": "SupportedProperty"
        },
        {
          "@id": ["schema:givenName"],
          "property": "sobrenome", "@type": "SupportedProperty"
        },
        {
          "@id": ["schema:image"],
          "property": "imagem", "@type": "Link"
        },
        {
          "@id": ["foaf:knows"],
          "property": "amigos", "@type": "SupportedCollection"
}
      ],
      "supportedOperations": [...]
    }.
    { demais classes semânticas ... }
  ]
}
```

Fonte: do **autor.**

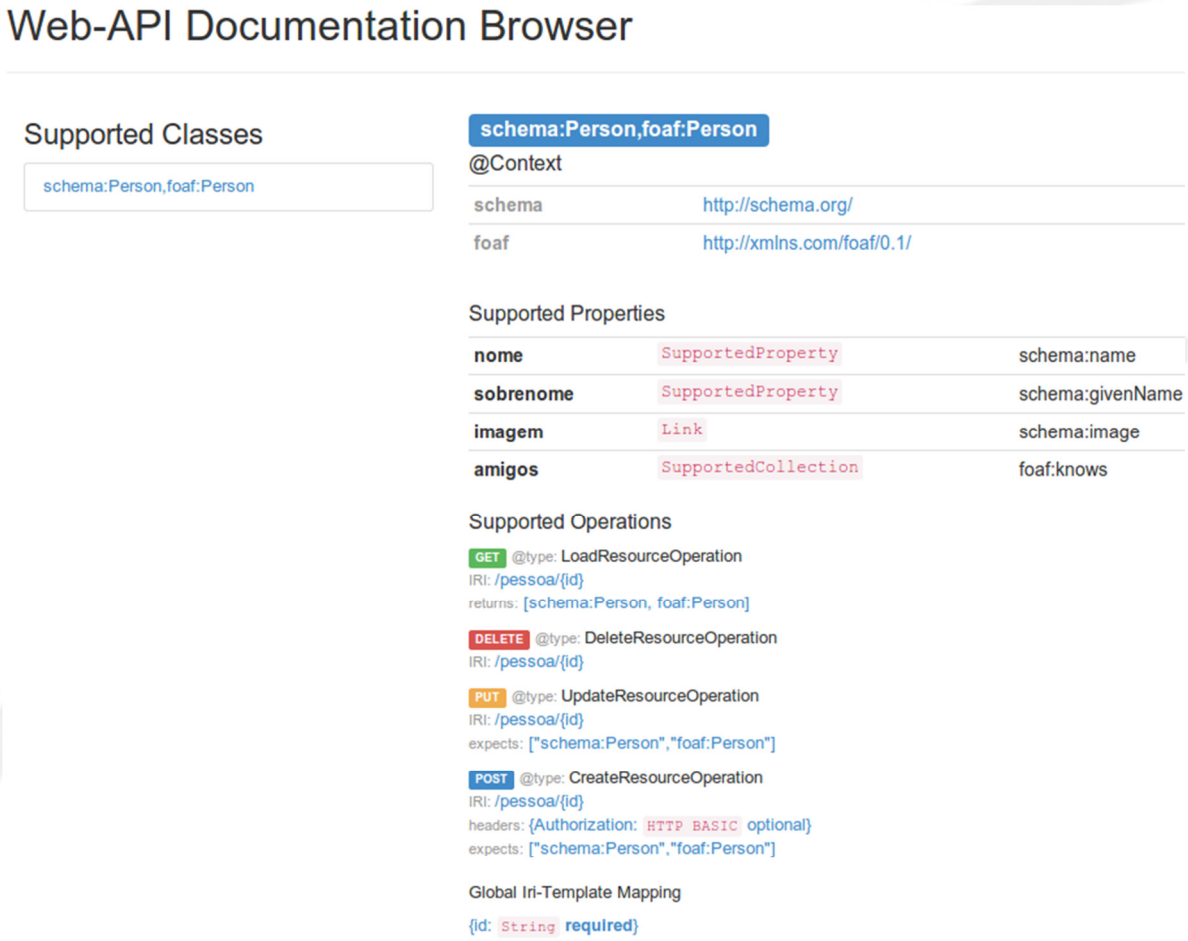
Figura 7– Exemplo de Documentação Hydra para Web APIs (supportedOperations)

```
  "supportedOperations": [
    {
      "method": "GET", "url": "/pessoa/{id}",
      "returns": ["schema:Person", "foaf:Person"],
      "@type": "LoadResourceOperation"
    },
    {
      "method": "DELETE", "url": "/pessoa/{id}",
      "@type": "DeleteResourceOperation",
      "headers": [{"header": "Authorization","@id": "HTTP BASIC"}]
     },
     {
      "method": "PUT", "url": "/pessoa/{id}",
      "expects": "["schema:Person","foaf:Person"]",
      "@type": "UpdateResourceOperation",
      "headers": [{"header": "Authorization","@id": "HTTP BASIC"}]
     },
     {
      "method": "POST", "url": "/pessoa/{id}",
      "expects": "["schema:Person","foaf:Person"]",
      "@type": "CreateResourceOperation",
      "headers": [{"header": "Authorization","@id": "HTTP BASIC"}]
}
]
```

Fonte: doautor.

ISBN 978-85-61115-09-8

Figura 8 – Exemplo de Documentação HTML JAX-SRS



Fonte: doautor.

3 Trabalhos Relacionados

Spring-HATEOAS é um *framework* destinado ao desenvolvimento de *Web Services* baseados em hipermídia na linguagem de programação Java. As representações de recursos são disponibilizadas no formato HAL (*Hypermedia Application Language*). Spring-HATEOAS apresenta o conceito de *Resource Representation Class*, que descreve apenas as propriedades da representação. As operações ficam separadas em classes *Resource Controller*, que manipulam as classes de representação e adicionam os *links* necessários. HAL é um formato para expressar controles hipermídia em documentos JSON e XML. Embora o formato JSON HAL permita representar controles hipermídia diretamente no documento JSON, os detalhes de execução necessitam de interpretação humana (SPRING, 2013).

Apache Isis é uma implementação Java da especificação RESTful Objects, extensível e personalizável, que faz uso de anotações para orientar a exposição de objetos de domínio. Utiliza um modelo arquitetural hexagonal, no qual os objetos de domínio assumem a posição central, enquanto o *framework* se responsabiliza pela persistência, segurança e apresentação. Pela sua natureza extensível e personalizável, o *framework* suporta diversas tecnologias, sendo que RESTful Objects é apenas uma delas. RESTful Objects é uma especificação para desenvolvimento de aplicações RESTful que permite o acesso aos objetos de domínio através de HTTP, resultando em representações no formato JSON. As classes de domínio podem

ISBN 978-85-61115-09-8

expor propriedades, operações e coleções que referenciam outras entidades. RESTful Objects permite representar dados e especificar tipos, além de vincular *links* com informações completas de execução (HAYWOOD, 2013).

JOHN; RAJASREE (2012)propõem um *framework* para descrição, descoberta e composição de serviços RESTful semânticos através de anotações semânticas na documentação da Web API. Embora enriquecer semanticamente a documentação de Web APIs proporcione maior integração com agentes autônomos, o cliente fica restrito à documentação como único guia para interação com a Web API.

SANCHEZ; OLIVEIRA; FORTES (2014) propõem o RestML, uma abordagem baseada no paradigma MDD (*Model Driven Development*) para facilitar o desenvolvimento de serviços Web. RestML baseia-se na construção de um modelo que resulta na geração automática de código fonte. Através de um *profile* UML, RestML permite que o modelo desenvolvido possa ser transformado em código fonte que obedece às definições arquiteturais REST. RestML propõe também um outro *profile* UML específico para a plataforma JavaEE.

WORDNIK (2014) apresenta um *framework* denominado Swagger, capaz de gerar a documentação de Web APIs. Permite que desenvolvedores e agentes de software compreendam como realizar a comunicação com serviços remotos. A documentação é gerada de forma automática através da análise do código fonte da Web API. Swagger é capaz de gerar a documentação de Web APIs desenvolvidas com várias tecnologias, como Java, Go, .Net, PHP, Scala, Ruby, dentre outras.

QMINO (2014) apresenta um *framework* denominado Miredot, capaz realizar a geração automática da documentação de Web APIs desenvolvidas em Java com o *framework* JAX-RS. A documentação é gerada através da análise do código fonte da aplicação e resulta em um documento HTML adequado para guiar desenvolvedores de aplicações clientes.

A Tabela 1 apresenta uma comparação entre os trabalhos relacionados e o *framework* propostoJAX-SRS. Dentre as características comparadas estão: suporte a controles hipermídia, descrição semântica, geração da documentação e o ponto de aplicação da proposta.

Sob o ponto de vista do suporte a controles hypermídia, Spring-HATEOAS proporciona suporte parcial, pois utiliza o formato de dados HAL que exige a intervenção humana para interpretar os detalhes das requisições. RestML, proprosto por SANCHEZ; OLIVEIRA; FORTES (2014), não oferece garantia de suporte a controles hipermídia, pois o resultado final é a geração automática de código para uma determinada plataforma. No caso da plataforma JavaEE, não existe suporte nativo para tais controles. O mesmo ocorre com as propostas de WORDNIK (2014) e QMINO (2014), que embora disponibilizem uma documentação que descreve as operações HTTP das Web APIs, não incorporam essas informações na forma de controles hipermídia diretamente sobre os recursos manipulados. O *framework* JAX-SRS incorpora as operações disponíveis para um determinado recurso no formato de controles hipermídia (operações Hydra) diretamente nas representações de recursos manipulados pela Web API.

Sob a perspectiva de descrição semântica, apenas a proposta de JOHN; RAJASREE, (2012) oferece algum tipo de suporte. Entretanto, esse suporte não corresponde diretamente à implementação da Web API, mas sobre a sua documentação gerada em formato HTML. De posse da documentação, o *framework* realiza anotações semanticas no documento, ampliando o entendimento da Web API para agentes de softwares e desenvolvedores.

ISBN 978-85-61115-09-8

Tabela 1 – Comparação entre os trabalhos relacionados

Trabalho Relacionado	Controles Hipermídia	Descrição Semântica	Geração da Documentação	Aplicado sobre
Spring-HATEOAS (SPRING, 2013)	Parcial	Não	Não	Código
Apache Isis (HAYWOOD, 2013)	Sim	Não	Não	Código
(JOHN; RAJASREE, 2012)	Sim	Sim	Não	Documentação
(SANCHEZ; OLIVEIRA; FORTES, 2014)	Não	Não	Parcial	Modelo
Swagger (WORDNIK, 2014)	Não	Não	Sim	Código
Miredot (QMINO, 2014)	Não	Não	Parcial	Código
JAX-SRS	Sim	Sim	Sim	Código

Fonte: do autor.

Dentre os trabalhos relacionados, apenas o *framework* Swagger realiza a geração e disponibilização de documentação de forma automática, da mesma forma que o *framework* JAX-SRS . Embora o *framework* Miredot realize a geração automática da documentação, a disponibilização não é direta, pois é realizada no momento de compilação da Web API e o resultado não é disponibilizado para consulta no contexto da aplicação. Em outras palavras, apenas o desenvolvedor tem a posse da documentação, enquanto a publicação deve ser feita através de um procedimento manual. Por esse motivo, Miredot é categorizado como parcial. A geração da documentação do RestML é considerada parcial pois a abordagem é baseada na construção de um modelo, que pode ser visto como uma forma de documentação da Web API. Entretanto, esse modelo pode apresentar detalhes de implementação que não são do interesse dos desenvolvedores de aplicações clientes.

Spring-HATEOAS, Apache Isis, Swagger, Miredot e JAX-SRS são aplicados diretamente no código fonte da Web API. A abordagem proposta por JOHN; RAJASREE (2012) tem objetivo de adicionar marcações semânticas na documentação pré-existente de uma Web API. Por fim, RestML utiliza a abordagem de modelagem UML que resulta na geração automática do código fonte da Web API. O ponto de aplicação do *framework* está mais relacionado com a natureza de cada projeto, e deve ser analisado individualmente.

Considerações Finais

Este artigo apresentou propostas para o desenvolvimento de Web APIs REST em representações com formato JSON. As propostas abordam questões de suporte ferramental para implementação de Web APIs REST semânticas com suporte a controles hipermídia. O suporte ferramental é apresentado através de um *framework* que permite ao desenvolvedor concentrar esforços no desenvolvimento do domínio do problema, sem a necessidade de esforço manual para adicionar informações semânticas aos recursos. Com a utilização do *framework* proposto, espera-se proporcionar maior produtividade e qualidade no desenvolvimento de Web APIs alinhadas com os princípios arquiteturais REST.

A descrição semântica de recursos Web através da associação com vocabulários controlados resulta em uma integração de dados e troca de informações muito mais flexível e rica, pois a estrutura e o significado dos dados são compartilhados em nível conceitual. O acoplamento entre Web API e seus clientes é reduzido, uma vez que a dependência concentra-se apenas no vocabulário compartilhado. Os mecanismos de descrição de operações Hydra reduzem ainda mais o acoplamento, pois oferecem as diretrizes necessárias para a execução de operações sobre os recursos Web. Dessa forma, as aplicações clientes não estão condicionadas ao conhecimento prévio dos detalhes da implementação da Web API.

O *framework* JAX-SRS oferece suporte ferramental para o desenvolvimento de Web APIs RESTful com descrição semântica e com controles hipermídia. O principal aspecto explorado pelo *framework* é a troca de documentos JSON constituídos não somente por

ISBN 978-85-61115-09-8

propriedades, mas também por controles hipermídia que permitem que aplicações clientes conheçam as operações disponíveis sobre um determinado recurso Web. A geração automática da documentação da Web API permite que aplicações clientes sejam programadas de forma a explorar com mais autonomia os recursos disponibilizados por Web APIs. Uma vez que os recursos Web e suas respectivas propriedades estão descritos semanticamente, uma aplicação cliente pode assumir uma postura exploratória em busca das informações desejadas e interagir com os recursos de acordo com a semântica de operações.

Como trabalhos futuros espera-seaprimorar detalhes sobre a composição de Web APIs, pois com a descrição semântica dos recursos Web possibilitada pelo JSON-LD, juntamente com o poder de representação de controles hipermídia proporcionado pelo Hydra, é possível que recursos sejam distribuídos entre diversas Web APIs, exigindo mecanismos de descoberta de recursos.

Referências

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. **The semantic web.** Scientific American, 2001. v. 284, n. 5, p. 34–43

FIELDING, R. T. **REST: Architectural Styles and the Design of Network-based Software Architectures.** Tese (Doctoral dissertation), University of California, Irvine, 2000. Disponível em: <http://www.ics.uci.edu/ fielding/pubs/dissertation/top.htm>.

HAYWOOD, D. **RRRADDD! Ridiculously Rapid Domain-Driven (and RESTful) Apps with Apache Isis.** 2013. Disponível em: < https://github.com/danhaywood/rrraddd-isis-131>.

JOHN, D.; RAJASREE, M. S. **A framework for the description, discovery and composition of restful semantic web services.** In: Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology . New York, NY, USA: ACM, 2012. (CCSEIT'12), p. 88.

LANTHALER, M. **Creating 3rd generation web apis with hydra.** International World Wide Web Conferences Steering Committee. ACM, 2013. p. 35–38.

LANTHALER, M.; GüTL, C. **On using json-ld to create evolvable restful services.** ALARCÓN, R.; PAUTASSO, C.; WILDE, E. (Ed.). WS-REST. ACM, 2012. p. 25–32.

ORACLE. **JAX-RS: Java API for RESTful Web Services – Public Review.** 2.ed. Oracle Corporation. 500 Oracle Parkway, Redwood Shores, CA 94065 USA, 2013.

QMINO. **Miredot – Reference Manual.**Qmino BVBA.2014. Disponível em: <http://www.miredot.com/docs/manual/>

RICHARDSON, L.; AMUNDSEN, M.; RUBY, S. **Restful Web Apis.** Oreilly &Associates Incorporated, 2013. ISBN 9781449358068.

SANCHEZ, R. V. V.; OLIVEIRA, R. R.; FORTES, R. P. M. **RestML: Modeling RESTful Web Services.** In: REST: Advanced Research Topics and Practical Applications. Springer New York, p. 125-143. 2014.

ISBN 978-85-61115-09-8

SPRING. **Spring Hateoas - Reference.** 2013. Disponível em: <http://projects.spring.io/spring-hateoas/>.

SVENONIUS, E. **Unanswered questions in the design of controlled vocabularies.** Journal of the American Society for Information Science, v. 37, n. 5, p. 331–340, set. 1986. ISSN 00028231.

WARNER, A. J. **A Taxonomy Primer.**Lexonomy, 2002. Disponível em: <https://www.ischool.utexas.edu/~i385e/readings/Warner-aTaxonomyPrimer.html>

WORDNIK. **The Swagger Specification.**Reverb Technologies, Inc. 2014. Disponível em: <https://github.com/wordnik/swagger-spec>

ISBN 978-85-61115-09-8

SMART CITIES BASEADAS EM *BIG DATA*: DESAFIOS E OPORTUNIDADES

Vinícius Barreto Klein
vinibk@gmail.com

José Leomar Todesco
tite@egc.ufsc.br

Resumo
Big Data é um fenômeno digital que vem sendo estudado por diversos pesquisadores e apontado como uma área de grandes desafios tecnológicos e oportunidades de negócio. *Smart cities* é um conceito que envolve certas características e habilidades que uma cidade deve possuir para elevar a qualidade de vida de seus habitantes. Conforme pesquisas realizadas, s*mart cities* podem utilizar-se de *bigdata* para atingir seus objetivos. Este trabalho visa apontar os desafios e oportunidades para a implantação de *smart cities* baseadas em *big data*. Para isso, realizou-se uma revisão integrativa que levantou o estado da arte destes temas. Foi identificado que projetos para implantar *smart cities* encontram diversos desafios atualmente, e os principais são: financeiro, complexidade tecnológica e estrutura política-administrativa das cidades. Para vencê-los, autores apontam três abordagens (ou modelos de negócio), onde as cidades devem gerenciar suas plataformas de acesso aos seus dados. Estas abordagens são chamadas de *App-Store, Google-Maps-Like* e *Open Data.* Esta última é criticada por seus autores porém, em nosso ponto de vista e contexto, cremos ser muito indicada. Ela oferece benefícios em demais domínios das *smart cities*, e não apenas no eixo tecnológico. Além disso, foi visto que o papel do desenvolvedor e da indústria de software são estratégicos. Eles devem conseguir trabalhar com uma nova geração de aplicativos,cuja principal funcionalidade é a habilidade de "sentir" o contexto de um usuário e adaptar-se em tempo real a ele, criando novos desafios e oportunidades para esta área.

Palavras-chave: *Big Data, Smart Cities,* Open Data, Desafios, Oportunidades.

Abstract
Big Data is a digital phenomenon that has been studied by many researchers and identified as an area of major technological challenges and business opportunities. Smart cities is a concept that involves certain characteristics and skills that a city should have to raise the quality of life of its inhabitants. According to the research conducted, smart cities can use big data to achieve their goals. This paper points out the challenges and opportunities for the deployment of smart cities based on big data. An integrative review was made to bring the state of art of these topics. It was identified that projects to implement smart cities have many challenges today, and the main ones are: financial, technological complexity and political-administrative structure of cities. To beat them, the authors suggest three approaches (or business models), where cities must manage their platforms to access their data. These approaches are called App-Store, Google-Maps-Like and Open Data. This last one is criticized by its authors but in our view and context, we believe it is very suitable. It offers benefits in other areas of smart cities, not just in the technological dimension. Furthermore, it was shown that the role of the developer and the software industry is strategic. They must be able to work with a new generation of applications, whose main feature is the ability to "sense" the context of a user and adapt itself in real time to it, creating new challenges and opportunities for this area.

Key Words: *Big Data, Smart Cities,* Open Data, Challenges, Opportunities.

ISBN 978-85-61115-09-8

Introdução

Segundo Fan e Bifet (2012), *big data* é um fenômeno que vem sendo estudado por diversos pesquisadores tanto da academia como da indústria, e, segundo eles, citando Parker (2012) e Gopalkrishnan (2012), representa uma área com diversos desafios futuros importantes, devido a natureza de seus dados: volumosos e evolutivos. Beyer e Laney (2012) afirma que *big data* exige novas formas de tratamento tecnológico. Isso se deve principalmente a seu alto volume e velocidade de produção, mas também por possuir diversos formatos e esquemas distintos (DAVENPORT, 2014; O'REILLY, 2012; BEYER e LANEY, 2012; FAN e BIFET, 2012). *Smart cities* é um conceito que envolve diversas características e habilidades que uma cidade deve possuir para se classificar como "s*mart*". (GIFFINGER, 2007). Uma destas habilidades envolve utilizar de TICs (tecnologias da informação e comunicação) para melhorar a qualidade de vida da população (CARAGLIU; BO; NIJKAMP, 2011). Conforme as pesquisas feitas para este trabalho, s*mart cities* podem utilizar-se de *big data* para atingir seus objetivos. DOBRE e XHAFA (2013) afirmam que são nas cidades que *big data* tem seu maior impacto, e as *smart cities* devem se basear neste fenômeno para melhorar a qualidade de vida de seus cidadãos. No entanto, implantar uma *smart city* a partir de *big data* é um processo que envolve certos desafios a serem considerados (VILAJOSANA et al., 2013). O objetivo deste trabalho é trazer o estado da arte de ambos estes conceitos, e em seguida, analisar criticamente soluções encontradas na literatura para construir *smart cities* a partir de *big data*, buscando entender e relatar seus desafios e oportunidades.

Procedimentos metodológicos

Para Botelho (2012), "a revisão da literatura é o primeiro passo para a construção do conhecimento científico, pois é por meio desse processo que são identificadas lacunas e oportunidades de pesquisa". Segundo a autora, citando MENDES, SILVEIRA, GALVÃO (2008), BENEFIELD (2003), POLIT e BECK (2006), a revisão integrativa é um método que possibilita sintetizar os materiais já publicados, contribuindo assim para a geração de novos conhecimentos, baseados nos conhecimentos já criados sobre os temas investigados.

Fazendo uso deste método, foi então analisado o estado da arte dos temas propostos e as soluções para o processo de implantação de *smart cities* baseadas em *big data.* Foram executados os seguintes passos:

- Busca na literatura: utilização de bases eletrônicas (amostra de coleta de dados textuais em bases científicas, em artigos publicados, livros de autores referenciados e definições de pesquisadores da academia e indústria). Foi focado na base de periódicos da CAPES, que reúne mais de 36 mil periódicos e 130 bases referenciais (CAPES, [20--]), além de outros meios, como livros, sites e relatórios técnicos;
- Discussão dos resultados: buscou-se analisar criticamente os resultados encontrados pelos autores dos trabalhos selecionados no passo anterior. Os resultados encontrados nesta pesquisa são relatados nas seguintes seções.
- Critérios para categorização dos estudos: foram buscados trabalhos que trouxessem o estado da arte dos tema pesquisados ou/e que estudassem soluções baseadas em *big data* para *smart cities*. Foram definidas palavras chaves que combinavam os temas estudados, ou os tratassem separadamente. Com auxílio do filtro por relevância disponível pela plataforma de periódicos da CAPES ([20--]), e através de uma leitura e análise dos títulos e resumos dos trabalhos encontrados, foi buscando identificar desafios e oportunidades nesta área. Os trabalhos que mais contribuíam para o problema proposto e que seguiam uma sequência lógica de complementação um do outro foram escolhidos para compor as referências desta pesquisa;
- Escolha e definição do tema: *smart cities, big data, smart cities based on big data,* conforme necessidade contextualizada na introdução deste trabalho.

ISBN 978-85-61115-09-8

1 *Big data*: causa e definições

No ano de 2011, o número médio diário de postagens (*tweets*) publicados no site Twitter foi de 200 milhões (TWITTER, 2011); no site Facebook.com, 30 bilhões de conteúdos são compartilhados mensalmente (MANYIKA et al., 2011); a cada um minuto são postados 72 horas de vídeo no Youtube (WOOLLASTON, 2013). Outro exemplo é a rede Walmart, que lida a cada hora com 1 milhão de transações em um banco de dados de aproximadamente 2,5 petabytes (10^{15} bytes) (ALLIANCE, 2014), o que corresponde aproximadamente a 170 vezes a quantidade de dados da Biblioteca do Congresso Americano (que possui em trono de 36,8 milhões de livros catalogados) (BETTINO, 2012). Segundo a IBM (2013), 90% dos dados existentes no mundo atual foram criados nos últimos dois anos. E ainda, em 2020, a tendência é que aproximadamente 26 bilhões de aparelhos sejam conectados a IoT (internet das coisas, do inglês, *internet of things*) (RIVERA; MEULEN, 2013).Na Figura 1 este cenário pode ser melhor visualizado.

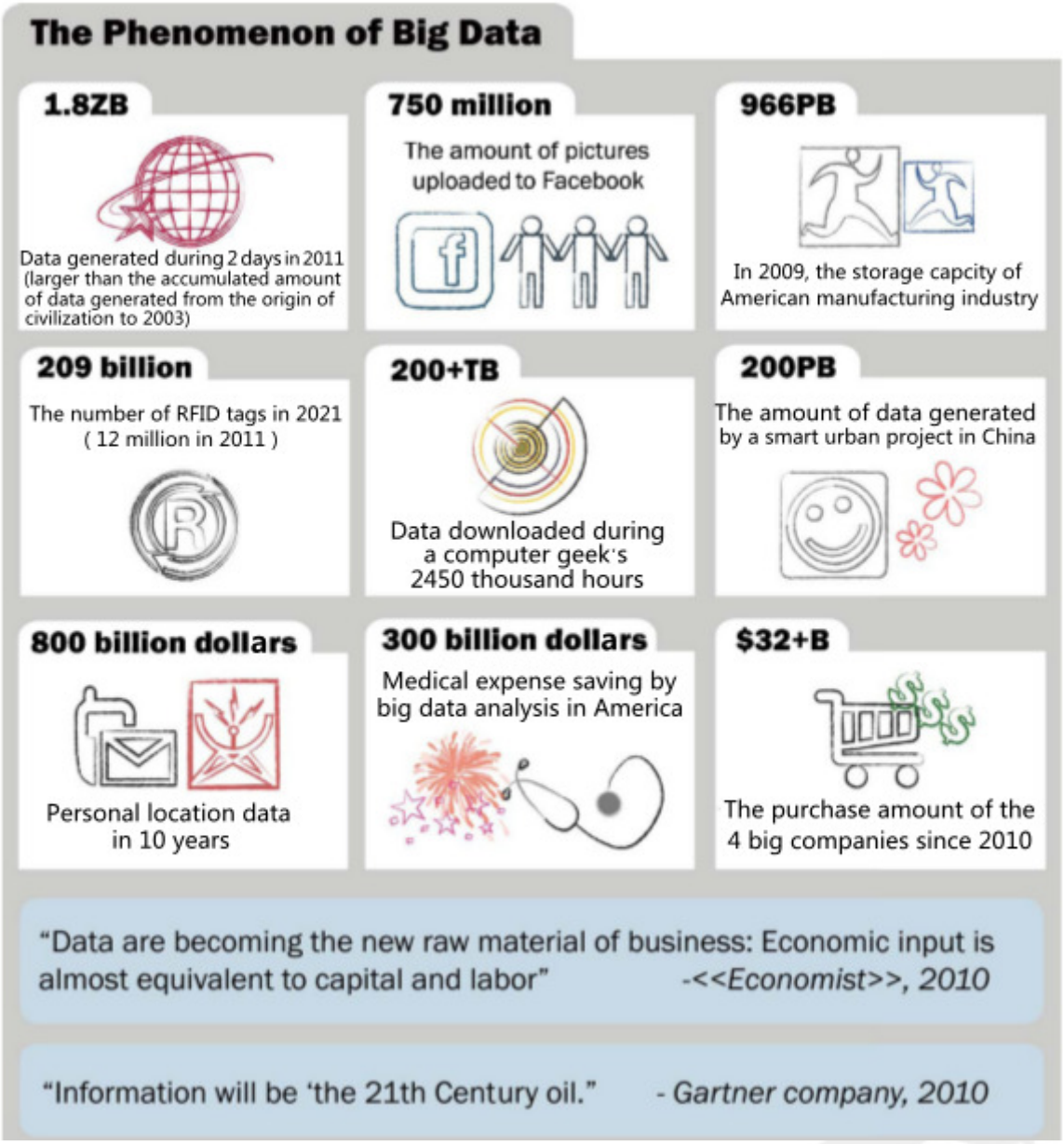
Toda esta quantidade de dados produzida corresponde ao fenômeno conhecido como *big data*. Este tema tem sido comumente relacionado a volumes de dados extremamente grandes, no entanto, existem diversas definições na literatura que evidenciam outras de suas características (CHEN; MAO; LIU, 2014). Fan e Bifet (2012), citando Laney (2001), definem o chamado modelo 3Vs:

- *Volume*: há hoje mais dados do que jamais houve (quantidade), porém, não há tantas ferramentas que possam processá-los;
- *Variety*: existem muitos tipos diferentes de dados, como texto, dados de sensores, audio, vídeos, imagens etc;
- *Velocity*: os dados chegam a fluxos contínuos, e necessita-se analisá-los em tempo real.

Nos anos seguintes a este modelo, segundo Chen, Mao e Liu (2014), citando Zikopoulos et al (2011) e Meijer (2011), empresas como a Gartner, Microsoft e IBM corroboraram e adaptaram este conceito em seus departamentos de pesquisa, com as seguintes definições: *volume*, corresponde a grandes massas de dados e altas escalas; *velocity*, refere-se a linha de tempo dos dados, que deve ser coletado e analisado rapidamente, para manter o valor comercial; e *variety*, que significa os vários formatos de dados, incluindo dados estruturados, semi e não estruturados. Seguindo esta linha, existem várias definições no meio acadêmico e de negócios.

ISBN 978-85-61115-09-8

Figura 1: O fenômeno big data



Fonte: CHEN; MAO; LIU, 2014

Abaixo, segue uma lista de algumas das definições encontradas na literatura, sob diferentes perspectivas e aspectos, que contribuem no entendimento deste fenômeno:

- "*Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures.*" (O'REILLY, 2012);
- "[...] *a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis* (GANTZ e REINSEL, 2011);
- "*Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.*"(GARTNER, [20--]).

ISBN 978-85-61115-09-8

- "*Big Data refers to data that is too big to fit on a single server, too unstructured to fit into a row-and-column database, or too continuously flowing to fit into a static data warehouse. While its size receives all the attention, the most difficult part of big data really involves its lack of structure*."(DAVENPORT, 2014).
- "*Big Data is a new term used to identify datasets that we can not manage with current methodologies or data mining software tools due to their large size and complexity.*" (FAN e BIFET, 2012).

Ainda definindo este conceito, alguns autores adicionam novas características, como *veracity*, que diz respeito ao fato de que estes dados devem ser confiáveis para que as organizações possam tomar suas decisões baseadas neles (IBM, 2013). Segundo Fan e Bifet (2012, tradução nossa), hoje em dia também são consideradas duas dimensões importantes em *big data*:

- *Variability*: corresponde aos diferentes esquemas dos dados coletados e como estes são interpretados;
- *Value*: valor de negócio, que possibilita às organizações tomarem decisões que antes (de *big data*) não era possível.

Como causas deste fenômeno podemos apontar o avanço do uso de TICs em diversos setores da sociedade (GROBELNIK, 2012). O uso de *cloud computing* e *da web 2.0,* assim como a alta inserção de *smartphones*, sensores e outros dispositivos produtores de dados na rotina das pessoas, são em termos gerais, características do avanço tecnológico que vivemos nos dias atuais. A era da informação e estes avanços com certeza contribuem para um maior volume, velocidade e variedade de dados produzidos. Grobelnik (2012), citando Manyika et al (2011) e Hillbert e Lopez (2011), destaca os principais fatores de causa do fenômeno *big data*:

- O aumento mundial da capacidade de armazenamento: considerando os aparelhos digitais e analógicos produzidos no período entre o ano de 2000 e 2007, estima-se que houve um crescimento de aproximadamente 50 exabytes (1 bilhão de gigabytes) para em torno de 290 exabytes;
- O aumento mundial da capacidade de processamento de dados: entre 2000 e 2007, esta capacidade aumentou de 10^{12} milhões de instruções por segundo para 6,5x 10^{12}. Os autores explicam que estes valores foram estimados através da multiplicação da quantidade de aparelhos instalados pela sua capacidade de processamento;
- A alta disponibilidade de dados: o mundo e as organizações possuem muito mais dados disponíveis. Segundo os autores, somente nos EUA, organizações de diversos setores, com mais de 1.000 empregados, possuem no mínimo 100 TB de dados armazenados.

Além disso, o aumento do uso das redes sociais e da internet das coisas contribuiu significativamente para o aumento desta disponibilidade.

2 *Smart cities*: definição de um modelo e suas características

O termo *smart cities* tem estado em evidência nos últimos anos. Segundo Townsend (2013, tradução nossa), *smart cities* são cidades que conseguem capturar dados, transformar em informação e adaptar suas ações em tempo real. Caragliu, Bo e Nijkamp (2011) descrevem uma *smart city* como uma cidade onde se investe em capital humano e social, utilizando-se da infraestrutura de tecnologia de informação para promover o crescimento econômico sustentável e a qualidade de vida. No entanto, nem todas as definições envolvem

ISBN 978-85-61115-09-8

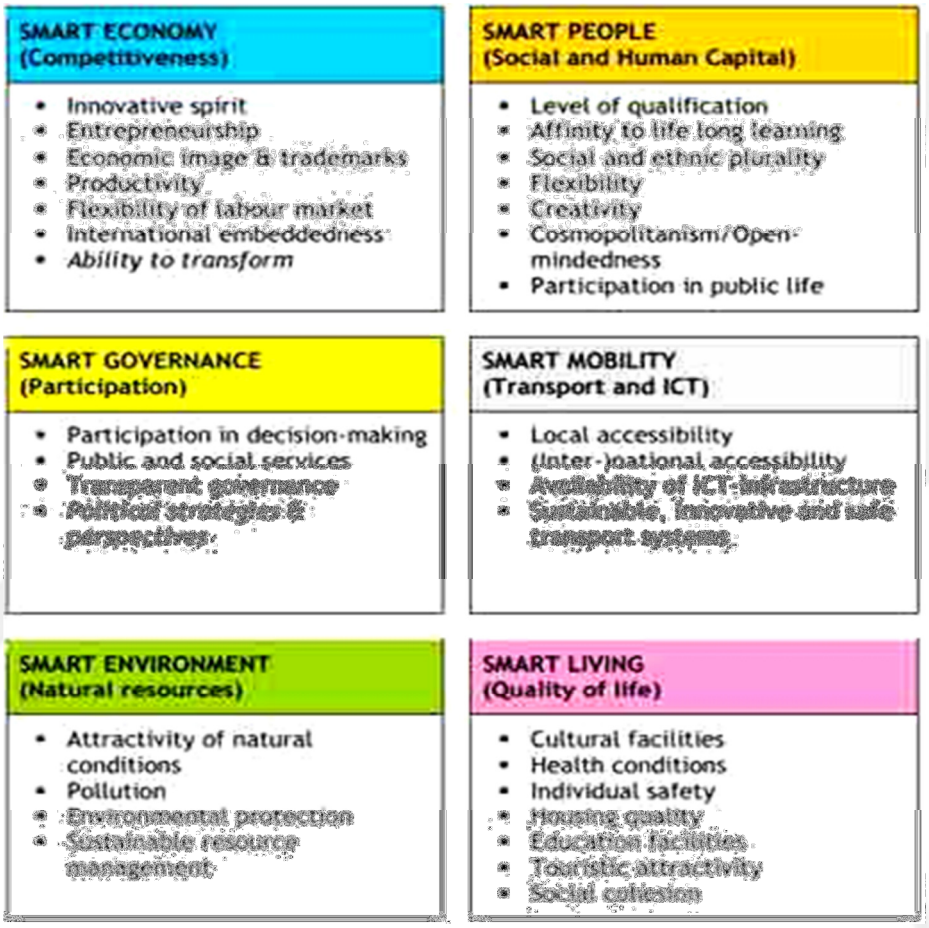
apenas o uso de TICs. Um trabalho que contribui bastante para o entendimento deste tema foi realizado pela Vienna University, que faz um estudo geral do estado da arte das definições existentes sobre *smart cities*. Os autores relatam que não há uma definição que considere apenas um único aspecto, mas que todas as definições encontradas na literatura envolvem um conjunto de características e habilidades que estas devem possuir, como desde ser um distrito de TI até possuir elevado nível de educação de seus habitantes (GIFFINGER, 2007). Com base nesta pesquisa eles então abstraem uma série de qualidades encontradas e criam um modelo de características genéricas encontradas nas definições, criando para cada característica um grupo de fatores com indicadores de performance. A Figura 2 apresenta este modelo.

Seguindo o modelo de Giffinger, com base principalmente em dados do Eurostat (escritório de estatística da união européia), em trabalhos de pesquisa do ministério de cultura e comunicação de Paris e do programa ESPON - *European Observation Network for Territorial Developement,* os autores então constroem um ranking que mostra a performance de um grupo de cidades de porte médio da Europa (eles escolhem este perfil de cidade justificando que as principais pesquisas urbanas focam em grandes metrópoles, mesmo com a maioria da população urbana vivendo em cidades de médio porte). Este *ranking* nos mostra como podem ser usados estes indicadores para medir a performance destas cidades, e nos dar uma ideia do que é preciso melhorar em uma determinada cidade por exemplo.

Ainda nesta linha, Neirotti et al (2014) fazem um interessante trabalho sobre as tendências atuais em *smart cities.* Os autores se baseiam em Giffinger (2007) e afirmam que os conceitos de *smart cities* não abrangem apenas inovações tecnológicas, mas também investimentos em capital humano e mudanças no modo de vida destas cidades. Nesta linha, eles classificam os domínios que uma *smart city* deve se preocupar em dois tipos: *hard e soft.* O primeiro diz respeito às áreas em que as TICs tem maior impacto, como em redes de energia, iluminação pública, fontes de energia renováveis, recursos hídricos, lixo, construções inteligentes (todas estas pensando na questão ambiental, evitando desperdícios de recursos através do uso de sensores e sistemas de informação adaptáveis ao contexto). Mobilidade urbana, saúde (utilizando TICs para oferecer melhor atendimento às pessoas) e segurança pública também entram nesta categoria. Já o segundo grupo (*soft*) é onde os sensores e informações adaptadas em tempo real tem menor impacto, pois são áreas que necessitam mais de ações (incentivos) que criem as condições necessárias para que elas se desenvolvam. São encaixadas neste grupo a economia (incentivo ao empreendedorismo, inovação, integração com mercado internacional), administração pública (digitalização da administração pública, *e-gov,* transparência digital para atrair o envolvimento dos cidadãos), educação (uso de TICs para criar mais acesso à educação, políticas educacionais), cultura (mais eventos, monumentos turísticos), inclusão social (diminuição de barreiras para melhorar a qualidade de idosos e pessoas com deficiência, por exemplo). E, logicamente, isso não quer dizer que o domínio *hard* não necessite destes tipos de incentivos.

ISBN 978-85-61115-09-8

Figura 2: Modelo de *smart cities* - características genéricas (em negrito) e seus fatores



Fonte: Giffinger, 2007

3 *Smart cities* baseadas em *big data:* desafios, abordagens e oportunidades

Com base nos conceitos apresentados é possível perceber que *smart cities* produzem e necessitam de *big data*. DOBRE e XHAFA (2013) afirmam que são nas cidades que *big data* tem seu maior impacto, e as *smart cities* devem se basear neste fenômeno para melhorar a qualidade de vida de seus cidadãos. No entanto, implantar uma *smart city* e ainda trabalhar com *big data* são atividades que exigem grandes desafios. Vilajosana et al (2013) propõem um modelo para se implantar uma *smart city* através do uso de *big data*. Neste trabalho, primeiramente os autores explicam que as grandes cidades atuais possuem necessidades reais de se tornarem *smart* devido a três fatores principais: mais da metade da população vive em ambientes urbanos (80% das cidades européias inclusive); os esforços para facilitar as condições de vida (e.g. mobilidade urbana) são enormes; e os efeitos da urbanização no clima global vêm se evidenciando cada vez mais. Eles afirmam que, observando estes fatores, governos tem procurado por soluções tecnológicas, e isso disparou o interesse de grandes fornecedores de serviços e produtos como, Cisco, IBM e HP, por exemplo, a produzirem suas próprias soluções. Porém, mesmo com ambas as partes interessadas (governo e indústria), os investimentos em grande escala ainda não estão acontecendo. Os autores então apontam os três principais desafios que impedem este desenvolvimento:

- **Desafio Político/Administrativo**: existe uma necessidade de mudanças nas estruturas político/administrativas das cidades. Por exemplo, uma prefeitura geralmente possui diversos setores e departamentos que lidam com as diversas áreas que uma *smart city* deve se

ISBN 978-85-61115-09-8

preocupar (meio-ambiente, tecnologias, turismo etc).No entanto, estes setores dificilmente funcionam de maneira integrada, principalmente do ponto de vista dos seus sistemas de informação, mas também nos processos de tomadas de decisões;

- **Desafio Tecnológico**: há certa complexidade de desenvolvimento nas soluções tecnológicas voltadas para as *smart cities/big data.* Estas soluções devem lidar com diversos tipos de dados, de diversas fontes, em diversos modelos e formatos e, em altos volumes e velocidade de produção (como visto na definição de *big data*);
- **Desafio financeiro**: é provavelmente o grande desafio a ser vencido pelas *smart cities.* A escassez de recursos financeiro (na Europa principalmente, devido a recente crise financeira) combinado ao alto custo das soluções tecnológicas para *big data* agravam a situação. Outro fator apontado é a ausência de claros modelos de negócio para se implantar uma *smart city*.

Para superar estes desafios os autores propõem as seguintes soluções:

- **Desafio Político**: é necessário criar departamentos de *smart cities* nos governos (assim como existem os departamentos de TI nas organizações, por exemplo). Essa ação diminuiria a complexidade na comunicação entre os *stakeholders* das *smart cities* e os diversos setores envolvidos (e atualmente não integrados), como saúde, meio ambiente, segurança, ciência e tecnologia por exemplo. Isso facilitaria também, segundo eles, o processo de tomada de decisão, que seria centralizado neste setor apenas, e a aproximação entre governo e iniciativa privada (e. g. indústria de software, pequenos e médios desenvolvedores, produtores de dispositivos para IoT), que possuiriam seu próprio canal;
- **Desafio Tecnológico**: é preciso implantar plataformas tecnológicas que consigam integrar os dados de diversos setores (com diferentes formatos e esquemas) e que consigam lidar com o grande volume de dados produzidos (oriundos desde sensores da *IoT* até de redes sociais e *open government data* por exemplo). Para este fim, uma solução é utilizar as já citadas grandes fornecedoras de serviços e produtos, que tem desenvolvido suas próprias soluções, ou desenvolver uma solução própria. Este desafio ainda é difícil de vencer, pois o custo de implantar ou desenvolver estas tecnologias é alto. Porém, as soluções para o desafio financeiro, listadas a seguir, podem auxiliar a trazer alívio financeiro para esta tarefa;
- **Desafio Financeiro**: este desafio traz uma oportunidade interessante. Segundo os autores, é possível atrair investimentos privados e filantrópicos nesta área (pelas questões ambientais ou de qualidade urbana, por exemplo) se for usada uma estratégia inteligente. É proposto por eles então um processo em três fases para atrair investimentos externos ao governo:
- **Fase 1:** focar em serviços que tragam retorno a curto prazo e que ofereçam informações com maior utilidade à vida dos cidadãos destas cidades. Aplicativos para segurança pública ou de trânsito, por exemplo, (como o recente Waze), que tragam melhoras rápidas e funcionalidades úteis. Se os governos executarem esta fase com sucesso, pode-se gerar mais rapidamente um fluxo de caixa para novos investimentos;
- **Fase 2:** focar em serviços com retorno financeiro a longo prazo. Espera-se que iniciativas privadas já sejam atraídas pelo fluxo de caixa e dados gerados na fase anterior. Aplicativos que coletem e façam a mineração de dados são um exemplo interessante para esta fase;
- **Fase 3:** disponibilização dos dados e alguns serviços gerados pelas duas fases anteriores, através de uma plataforma para acesso. Fazendo isso, as *smart cities* podem gerar novas receitas, cobrando (ou não) pelo uso dos dados. Segundo os autores, o acesso a esta plataforma pode ser gerenciado pelas seguintes abordagens ou modelos de negócio:
- **Abordagem *App-Store-Like***: os desenvolvedores se inscrevem e acessam os dados através o uso de APIs, para que estes criem seus próprios aplicativos. Pode ser inclusive

ISBN 978-85-61115-09-8

cobrada uma taxa para este acesso. Os aplicativos desenvolvidos podem ser depois disponibilizados em repositórios de comercialização de aplicativos para dispositivos móveis, como a Google Play, para Android ou outras *App Stores*, por exemplo*;*

- **Abordagem *Google-Maps-Like***: serviços que necessitem de dados com alta granularidade, confiabilidade e veracidade, exemplificando, como em aplicativos de trânsito, podem ter uma taxa cobrada conforme o nível de detalhes ou frequência de acesso aos dados. Um exemplo deste tipo de cobrança é a API Google *Prediction*, que cobra conforme a frequência de consulta aos dados aumenta;
- **Abordagem *Open-Data*** (assim chamado pelos autores): as cidades podem fornecer acesso aos dados sem cobrança de taxas. Os autores, porém crêem que esta abordagem não é financeiramente viável e precisaria de subsídio financeiro governamental.

Estes três modelos de negócio e o processo trifásico analisados reúnem desafios e oportunidades interessantes para projetos desta natureza, além de apontarem o papel dos desenvolvedores e indústria de software como estratégico. Logo, é necessário que este setor conheça quais características seus produtos devem possuir, para conseguir entrar nesta gama de oportunidades criada pelas *smart cities.* Uma oportunidade interessante surge das características existentes nos atuais dispositivos móveis combinados à IoT. Sabe-se que a internet das coisas é uma importante fonte de *big data* (CHEN; MAO; LIU, 2014) e é uma parte essencial da infraestrutura das *smart cities*. Logo, além dos tradicionais sensores das *smart cities* (em redes de água, calculando a vazão, por exemplo), os atuais dispositivos móveis (*smartphones e tablets,* por exemplo) podem atuar também como sensores, coletando e gerando dados. Eles podem produzir informações sobre o contexto de um usuário em uma determinada situação, possibilitando a criação de uma nova geração de aplicativos baseados em *big data* (DOBRE e XHAFA, 2013). Imagine por exemplo, uma aplicação que consiga coletar a localização atual de um determinado usuário, através do GPS de seu dispositivo móvel, e combinar com suas preferências (gostos por certos tipos de produtos). Este aplicativo conseguiria enviar informações ao seu usuário sobre onde encontrar lojas de interesse mais próximos ao local em que ele estivesse em um determinado momento. Esta ideia está relacionada ao conceito de computação sensível ou perceptível ao contexto, onde os computadores podem perceber uma determinada situação e reagir a ela (ROBLES e KIM, 2010). Nesta linha, segundo Gartner (2009), captar e reagir ao contexto de um usuário será tão influente em *apps* para dispositivos móveis quanto as máquinas de busca são para a web. Estes aplicativos representam a ideia de que cidades inteligentes são cidades que agem e reagem aos acontecimentos de seu cotidiano (NEIROTTI, et al. 2014). Isso direciona e dispara uma gama de oportunidades de desenvolvimento de aplicativos para *smart cities*. Dobre e Xhafa (2013) chamam este tipo de aplicativo como a nova geração de aplicações, e expõem quais características devem ser consideradas como requisitos por elas:

- Mobilidade e localização: as aplicações devem auxiliar o usuário a encontrar informações sobre lugares próximos a ele;
- Proximidade: a quantidade de dados em determinadas situações pode ser alta demais, e os dispositivos móveis tem capacidade de armazenamento e processamento limitada. É preciso que se filtrem os dados, trazendo apenas dados geograficamente mais próximos ao usuário;
- Garantia *Real-time:* as informações precisam ser atuais. O usuário não deve receber informações obsoletas (notícias sobre acontecimentos do mês passado, por exemplo);
- Suporte para erros de comunicação: nenhuma aplicação para *smart cities* deve assumir que o usuário está sempre conectado à Internet, ou com uma conexão que nunca caia;

ISBN 978-85-61115-09-8

- Descoberta de novas fontes: as aplicações devem conseguir descobrir novas fontes de dados, como outros sensores e serviços externos. Isso aumentaria a possibilidade de informações e funcionalidades úteis (interação com prédios inteligentes poderiam abrir uma porta ou acionar o elevador, por exemplo);
- Escalabilidade: devem ser escaláveis e possuir histórico de armazenamento, para permitir futuras minerações. Isso permite futuras aplicações de mineração de dados muito interessantes. O problema aqui seria a baixa capacidade de armazenamento dos dispositivos, mas que poderia ser resolvido com *cloud computing*.

As características de contexto acima apresentadas, somadas aos modelos de negócio analisados, auxiliam no caminho para implantar cidades inteligentes, trazendo grandes desafios e oportunidades a serem vencidos pelas partes envolvidas.

Considerações finais

Com base nos estudos expostos é possível perceber quais são os principais desafios encontrados pelas *smart cities* e as oportunidades geradas por elas. Além de precisarem atuar em diversos eixos (*smart econonomy, smart people, smart governance, smart environment, smart living, smart mobility*), as *smart cities* necessitam trabalhar com *big data.* Isso gera um grande desafio tecnológico e alto custo de implantação, além dos desafios político/administrativos que precisam ser vencidos. No entanto, é possível se guiar pelo processo e modelos de negócio analisados aqui para se implantar uma *smart city* a partir do uso de *big data*, e de forma auto-sustentável financeiramente, o que representa uma grande oportunidade para os administradores destas cidades. Consequentemente, isso gera também grandes oportunidades para os desenvolvedores de software e fornecedores de produtos e serviços de TI, que possuem papel estratégico neste processo. Eles devem estar atentos à intersecção entre o conceito de computação sensível ao contexto e *smart cities*, traduzida no conjunto de características exigidas por esta nova geração de aplicativos, que precisam considerar contexto de um usuário e interagir com ele (DOBRE e XHAFA, 2013).

Ainda sobre os modelos analisados, a abordagem chamada de *Open Data* (VILAJOSANA et al, 2013) é apresentada como não aconselhada do ponto de vista financeiro, pois não é cobrado pelos acesso aos dados produzidos pelas plataformas de *smart cities/big data* das cidades, o que seria inviável financeiramente segundo seus autores. Por outro lado, o movimento *Open Data,* que torna os dados públicos abertos, traz diversos benefícios para a sociedade, como transparência e controle social, participação popular, geração de novos serviços, geração de conhecimento e inovação, por exemplo, (FOUNDATION, [20--]), além de estar em conformidade com a lei brasileira de acesso a informação (lei 12.527, de novembro de 2011). Nesta linha, analisando o modelo de características genéricas de uma *smart city*, estas vantagens podem ser associadas às qualidades *Smart Governance* (transparência), *Smart People* (participação na vida pública, geração de conhecimento) e *Smart Economy* (empreendedorismo) respectivamente. A Tabela 1 mostra os resultados que podem ser encontrados em cada domínio de uma *smart city* quando aplicada cada abordagem distinta. Podemos perceber nela dois fatos: o impacto (*hard* e *soft)* das tecnologias baseadas em *big data* é menor em alguns eixos (Neirotti et al 2014), e principalmente as grandes contribuições do modelo *Open Data* aqui citadas (em negrito na tabela).

ISBN 978-85-61115-09-8

Tabela 1 – Impactos das abordagens de uso de *big data* para *smart cities*

Abordagem/Eixo e Impacto	*Smart* Environment	*Smart Mobility*	*Smart Living*	*Smart Governance*	*Smart Econonomy*	*Smart People*
Abordagem *App-store-Like-Model*	Potencialmente Alto	Potencialmente Alto	Potencialmente Alto	Potencial Alto/Transparência limitada	Impacto potencialmente alto no fator Empreendedorismo	Menor Impacto
Abordagem *Google-Maps-Like-Model*	Potencialmente Alto	Potencialmente Alto	Potencialmente Alto	Potencial Alto/ Transparência limitada	Impacto potencialmente alto no fator Empreendedorismo	Menor Impacto
Abordagem *Open-Data-Like-Model*	Potencialmente Alto	Potencialmente Alto	Potencialmente Alto	Potencial Alto/**Aumenta a Transparência / em conformidade com legislação**	Impacto potencialmente alto no fator Empreendedorismo	Maior **Impacto no fator participação em vida pública** / Geração de conhecimento

Fonte: Os autores

Cremos que, apesar desta abordagem ser apontada como desvantajosa do ponto de vista financeiro, ela apenas possui menor retorno financeiro direto, porém pode trazer retorno financeiro de formas indiretas e em outros formatos. Primeiramente, este problema de retorno direto pode ser encarado como um desafio que também necessita de estratégias criativas de solução, o que é um interessante tema para trabalhos futuros. Inspirado nestas abordagens, um exemplo seria talvez uma combinação delas, criando um modelo de negócio que cobre taxas apenas de aplicativos que usem os dados para fins comerciais (algo como um *commercial open data,* utilizando *copyleft* e baseando-se em estudos sobre o valor comercial de *open data,* por exemplo). Isto pode ser estudado em trabalhos futuros. O grande benefício é que, diferente das outras duas abordagens, nela a transparência pública seria forçada, pois não apenas desenvolvedores acessariam os dados. Ela também poderia auxiliar no eixo *Smart Governance,* no fator "participação em vida pública" (*Smart People*), uma vez que os dados devem ser disponibilizados de forma pública, fazendo com que a população no mínimo fique ciente das transações governamentais, possivelmente motivando-a a exercitar mais a sua cidadania. Este eixo também seria privilegiado pela geração de conhecimento, podendo aumentar o nível de seus habitantes, que teriam mais possibilidades de construção de conhecimentos com base nestes dados (uso por instituições de ensino, por exemplo). Além de tudo, de maneira geral, dados podem ser usados como importantes fontes de desenvolvimento socioeconômico (HILBERT, 2013). A grande contribuição do modelo de negócio *Open Data* é que ele pode trazer diversos benefícios que vão além do retorno financeiro imediato, representando uma oportunidade de desenvolvimento de cidades inteligentes em diversos domínios, tanto tecnológicos como humanos e sociais.

ISBN 978-85-61115-09-8

REFERÊNCIAS

ALLIANCE, University. **What is Big Data?** Disponível em: <http://www.villanovau.com/resources/bi/what-is-big-data/#.VBZnxZ_7G01>. Acesso em: 20 ago. 2014.

BEYER, Mark A.; LANEY, Douglas. **The Importance of 'Big Data': A Definition**. 2012. Disponível em: <https://www.gartner.com/doc/2057415/importance-big-data-definition>. Acesso em: 27 set. 2014.

BETTINO, Larry A., **Transforming Big Data Challenges Into Opportunities**. Information Management. Extraído de<http://bit.ly/PWKQtq>. Acessado em 18/08/2014.

BOTELHO, Louise de Lira Roedel. **Aprendizagem Gerencial na Mudança em uma Organização Intensiva em Conhecimento**. 2012. 262 f. Tese (Doutorado) - Curso de Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Departamento de Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2012. Disponível em: <http://btd.egc.ufsc.br/wp-content/uploads/2013/01/LOUISEBOTELHO.pdf>. Acesso em: 14 ago. 2014.

BRASIL. **Lei no 12.527**, de 18 de novembro de 2011. In: Diário Oficial da República Federativa do Brasil, Brasília, DF, 18 nov. 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em 14/09/2014.

CAPES, Portal de Periódicos da. **Acervo**. [20--]. Disponível em: <http://periodicos.capes.gov.br/index.php?option=com_pcollection&Itemid=104>. Acesso em: 29 set. 2014.

CARAGLIU, Andrea; BO, Chiara del; NIJKAMP, Peter. Smart Cities in Europe. **Journal Of Urban Technology**. Londres, p. 65-82. 10 ago. 2011. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/.VBZNMx_Hk8o#.VBZPMZ_7G00>. Acesso em: 11 jun. 2014.

CHEN, Min; MAO, Shiwen; LIU, Yunhao. Big Data: A Survey. **Springer**, Estados Unidos, v. 2,n.19,p.171-209,Abr.2014. Disponível em: <http://link.springer.com/article/10.1007%2Fs11036-013-0489-0>. Acesso em: 13 set. 2014.

DAVENPORT, Thomas. **Big Data at Work:** Dispelling the Myths, Uncovering the Opportunities. Boston: Harvard Business Publishing Corporation, 2014. 229 p.

DOBRE, Ciprian; XHAFA, Fatos. Intelligent services for Big Data science. **Future Generation Computer Systems**. Amsterdam, p. 267-281. jul. 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167739X13001593>. Acesso em: 05 set. 2014.

FAN, Wei; BIFET, Albert. Mining Big Data: Current Status, and Forecast to the Future. **SIGKDD Explorations**, China, v. 2, n. 14, p.1-5, mar. 2012. Disponível em: <http://www.kdd.org/sites/default/files/issues/14-2-2012-12/V14-02-01-Fan.pdf>. Acesso em: 15 set. 2014.

ISBN 978-85-61115-09-8

FOUNDATION, Open Knowledge. **Por que Abrir Dados?** [20--]. Disponível em: <http://opendatahandbook.org/pt_BR/why-open-data/index.html#why-open-data>. Acesso em: 20 ago. 2014.

GANTZ, John; REINSEL, David. Extracting Value from Chaos. **IDC Iview**, Framingham, v. 01, n. 17, p.1-12, jun. 2011. Disponível em: <http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf>. Acesso em: 13 set. 2014.

GARTNER. **Big Data**. IT Glossary [20--]. Disponível em <http://www.gartner.com/it-glossary/big-data/>. Acessado em 19/08/2014.

GARTNER, Inc.. **Gartner Says Context-Aware Computing Will Provide Significant Competitive Advantage**. 2009. Disponível em: <http://www.gartner.com/newsroom/id/1190313>. Acesso em: 19 set. 2014.

GIFFINGER, Rudolf et al. **Smart cities Ranking of European medium-sized cities**. Vienna: Vienna University Of Technology, 2007. 28 p. Disponível em: <http://www.smart-cities.eu/download/smart_cities_final_report.pdf>. Acesso em: 06 jun. 2014.

GROBELNIK, Marko. **Big Data Tutorial**. Kalamaki: Jožef Stefan Institute, 2012. Color. Disponível em: <http://videolectures.net/eswc2012_grobelnik_big_data/>. Acesso em: 22 jul. 2014.

HILBERT, Martin; LÓPEZ, Priscila. **The World's Technological Capacity to Store, Communicate, and Compute Information**. Science, Nova York, v. 60, n. 332, p.60-65, jun. 2012.

HILBERT, Martin. Big Data for Development: From Information- to Knowledge Societies. **Social Science Electronic Publishing**, Rochester, jan. 2013. Disponível em: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2205145>. Acesso em: 28 set. 2014.

IBM, Software. **The IBM big data platform**. Somers: IBM Software, 2013. 4 p. Disponível em: <http://public.dhe.ibm.com/common/ssi/ecm/en/imb14135usen/IMB14135USEN.PDF>. Acesso em: 15 ago. 2014.

MANYIKA, James et al. **Big Data: The next frontier for innovation, competition, and productivity**. Nova York: Mckinsey Global Institute, 2011. 20 p. Disponível em: <http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>. Acesso em: 18 jun. 2014.

NEIROTTI, Paolo et al. Current trends in Smart City initiatives: Some stylised facts. **Cities: the international journal of urban policy and planning**. Amsterdã, p. 25-36. jun. 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0264275113001935>. Acesso em: 15 set. 2014.

O'REILLY, Media. **Big Data Now**. Sebastopol: O'reilly Media, 2012. 119 p.

RIVERA, Janessa; MEULEN, Rob van Der. **Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020**. 2013. Disponível em: <http://www.gartner.com/newsroom/id/2636073>. Acesso em: 02 jul. 2014.

ISBN 978-85-61115-09-8

ROBLES, Rosslin John; KIM, Tai-hoon. Review: Context Aware Tools for Smart Home Development. **International Journal Of Smart Home**. Daegu, p. 1-12. jan. 2010. Disponível em: <http://www.sersc.org/journals/IJSH/vol4_no1_2010/1.pdf>. Acesso em: 23 set. 2014.

SAS, Statistical Analysis System -. **O que é big data**. 2011. Disponível em: <http://www.sas.com/offices/latinamerica/brazil/solucoes/bigdata/>. Acesso em: 18 ago. 2014.

TOWNSEND,A.**Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia.** W. W. Norton & Company.2013.

TWITTER, Engineering. **200 million Tweets per day**. 2011. Disponível em: <https://blog.twitter.com/2011/200-million-tweets-day>. Acesso em: 18 ago. 14.

THE OFFICIAL TWITTER BLOG. **200 million Tweets per day**. Extraído de: <https://blog.twitter.com/2011/200-million-tweets-day>. Acessado em 18/08/2014.

VILAJOSANA, Ignasi et al. Bootstrapping smart cities through a self-sustainable model based on big data flows. **IEEE Communications Magazine**, Estados Unidos, v. 51, n. 6, p.128-134, jun. 2013. Disponível em: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&tp;=&arnumber=6525605&queryText;=vilajosana+big+data+bootstraping>. Acesso em: 12 jul. 2014.

WOOLLASTON, Victoria. **Revealed, what happens in just ONE minute on the internet: 216,000 photos posted, 278,000 Tweets and 1.8m Facebook likes**. 2013. Disponível em: <http://www.dailymail.co.uk/sciencetech/article-2381188/Revealed-happens-just-ONE-minute-internet-216-000-photos-posted-278-000-Tweets-1-8m-Facebook-likes.html>. Acesso em: 12 ago. 2014.

ISBN 978-85-61115-09-8

UMA ABORDAGEM PARA A PUBLICAÇÃO DE DADOS LIGADOS OBTIDOS A PARTIR DE BASES DE DADOS RELACIONAIS

Clayton Martins Pereira
clayton.martins@inpe.br

José Maria Parente de Oliveira
parente@ita.br

Resumo
Este trabalho tem como objetivo apresentar uma proposta de abordagem para facilitar e automatizar a publicação, na *Web* Semântica, de dados abertos ligados obtidos a partir de bases de dados relacionais (BDRs), por meio da integração entre as diversas ferramentas de software aplicadas neste processo. Oferece ainda uma nova ferramenta que possibilita a customização, de forma semi-automática, do arquivo de mapeamento entre a BDR e o modelo de dados *RDF*, a fim de incorporar a este uma ontologia de domínio fornecida pelo usuário. A abordagem proposta, chamada de *RDB2LOD*, apresenta como diferencial a automatização, por meio de um aplicativo (interface gráfica), das ferramentas utilizadas para a geração e customização deste arquivo de mapeamento, e para a visualização e consulta dos dados abertos ligados obtidos a partir dele. Conclui-se que a abordagem *RDB2LOD* permitirá a publicação na *Web*, em grande escala, de dados ligados obtidos a partir de bases de dados relacionais, de forma a possibilitar amplo acesso a muitos usuários, o que hoje ocorre ainda de forma muito tímida por meio de algumas iniciativas.

Palavras-chave: Dados Ligados. *Web* Semântica. Ontologias. Bases de Dados Relacionais.

Abstract
This work aims to present a proposal of a linked open data approach to facilitate and automate publishing, in the Semantic Web, obtained from relational database (RDB), through the integration between the various software tools used in this process. It also offers a new tool that enables the semi-automatic customization of the mapping among RDB and the RDF data model in order to incorporate a domain ontology supplied by the user. The proposed approach, called RDB2LOD, presents as differential the automation through an application (GUI) of the tools used for generating and customizing of this mapping file, and to view and query the linked open data obtained from it. It is concluded that the RDB2LOD approach will enable web publishing of linked data from relational databases, so as to allow well as a broad access to many users, which still occurs today in a very shy way through some initiatives.

Key Words: Linked data. Semantic web. Ontology. Relational databases.

1 Introdução

Atualmente, há uma vasta quantidade de dados armazenados em Bases de Dados Relacionais (BDRs) (SAHOO et al., 2009).Este é o modelo mais comumente utilizado e constitui o núcleo da maioria dos sistemas de tecnologia da informação (TI) hoje em uso (CHOI et al., 2010).Isso se deve à maturidade e eficiência de sua forma de armazenagem e consulta a dados, além de sua alta confiabilidade e escalabilidade (LING; ZHOU, 2010). Estima-se que 70% dos sítios *Web* são alimentados por dados mantidos em BDRs(SEQUEDA; ARENAS; MIRANKER, 2012).

ISBN 978-85-61115-09-8

Por outro lado, busca-se uma evolução da atual "*Web* de documentos" para uma "*Web* de dados",por meio de uma "*Web* Semântica",na qual as informações possam ser disponibilizadas em formato aberto (bruto) e apresentem significados bem definidos, ou seja, exatamente de acordo com o contexto ou domínio de conhecimento ao qual estas sejam aplicadas (semântica), de forma a possibilitar que computadores sejam capazes de processá-las e entendê-las automaticamente (BERNERS-LEE; HENDLER; LASSILA, 2001). Isso permitirá que tais informações sejam compartilhadas a partir de diferentes fontes, além de habilitar seu uso em diferentes contextos por diversas aplicações e agentes de software (LV; MA, 2008). Desta forma, o sucesso da *Web* Semântica depende da criação em massa de dados em formato aberto (LING; ZHOU, 2010).

Neste sentido, disponibilizar as grandes quantidades de dados armazenados em BDRs em formato aberto, por meio da*Web* Semântica,pode permitir que diferentes bases sejam unificadas, bem como seus dados sejam associados a outros dados estáticos ou outras fontes de dados (dados ligados), de forma a se obter um "ganho de informação". Além disso, promoveria ainda a interoperabilidade entre os sistemas de informação existentes e a formação de uma "rede de dados interpretáveis universalmente", a qual poderia ser usada como meio para a construção de aplicações orientadas a dados (LV; MA, 2008).

Diversas ferramentas de *software* têm sido criadas com este propósito, porém verifica-se uma carência de aplicações que promovam a integração e a automatização da operação destas ferramentas, bem como permitam a incorporação de ontologias de domínio ao processo de mapeamento entre a Base de Dados Relacional (BDR) e o formato de dados abertos.

Este artigo tem como objetivo apresentar uma proposta de abordagem para facilitar e automatizar a publicação de dados ligados obtidos a partir de BDRs. Nela são oferecidas uma aplicação com a finalidade de integrar ferramentas de software existentes, e uma nova ferramenta paraa customização semi-automática do arquivo de mapeamento entre a BDR e o formato de dados abertos, de forma a possibilitarque seja nele incorporada uma ontologia de domínio fornecida pelo usuário.

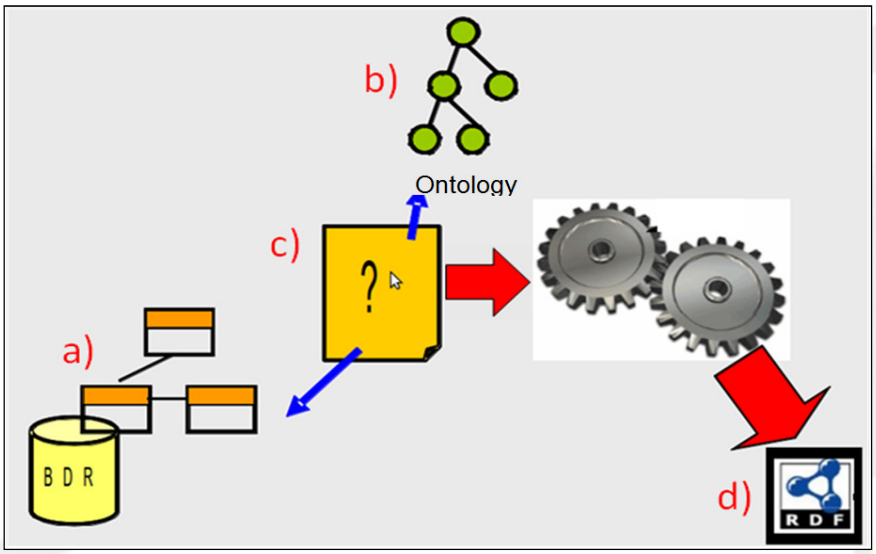
2Referencial Teórico e Trabalhos Relacionados

A *Web* atual, de uma forma geral, tem seu conteúdo destinado à leitura humana (foco no nível de apresentação dos dados), o que impossibilita sua manipulação por computadores. Com a *Web* Semântica, pretende-se que as informações disponibilizadas na *Web* estejam em um formato estruturado, com significados bem definidos, habilitando os computadores a processar e entender automaticamente estes dados (BERNERS-LEE; HENDLER; LASSILA, 2001). Isso promoverá uma mudança da atual "*Web* de documentos" para uma "*Web* de dados" (ANTONIOU; HARMELEN, 2008). Os principais padrões e artefatos atualmente disponíveis para a *Web* Semântica são o modelo de dados *RDF*, as ontologias (lista de termos e relações que descrevem formalmente um domínio de conhecimento), os dados abertos ligados, e a linguagem de consulta *SPARQL*.

O processo de conversão dos dados de uma BDR para um *datasetRDF*, de forma que possa ser publicado na *Web* Semântica no formato de dados abertos ligados, está ilustrado na Figura 1.

ISBN 978-85-61115-09-8

Figura 1 – Processo de conversão de uma BDR para um *dataset RDF*



Fonte: Autor

O processo apresentado na Figura 1 pode ser entendido da seguinte forma: a partir da BDR (a), e com base em uma ontologia (b), é gerado (de forma manual ou automática) um arquivo de mapeamento (c), escrito em uma linguagem específica, no qual são relacionados os mapeamentos entre as tabelas da BDR e as classes da ontologia, e entre as colunas da BDR e as propriedades da ontologia; e, com base no arquivo de mapeamento gerado é efetuada, por meio de ferramentas específicas, a transformação de cada linha da BDR em triplas *RDF* (d).

O mapeamento entre BDR e *RDF* pode atualmente ser efetuado por meio de duas técnicas: o mapeamento estático ou o mapeamento dinâmico. O mapeamento estático, também chamado de *RDF dump*, é uma técnica baseada em *ETL* (*Extract, Transform and Load*), onde um repositório *RDF* é criado a partir da conversão de todas as tabelas e colunas de uma BDR, com base em um arquivo de mapeamento. Já o mapeamento dinâmico efetua a conversão das tabelas e colunas de uma BDR para o formato de triplas *RDF* sob demanda (*on-line*), ou seja, de acordo com as respostas a serem retornadas às consultas elaboradas pelo usuário ou pelas ferramentas de publicação de dados abertos ligados (ZHOU, 2010).

A seguir é apresentada uma análise das ferramentas para publicação de dados abertos ligados propostas em alguns trabalhos relacionados e, ao final, é mostrado um comparativo entre estas e a ferramenta desenvolvida na abordagem proposta neste artigo.

A ferramenta *DB2OWL* (CULLOT; GHAWI; Y'ETONGNON, 2007) tem por objetivo converter automaticamente um esquema de BDR em uma ontologia *OWL*, por meio do mapeamento de tabelas em classes e de colunas em propriedades. Já as subclasses e os *object properties* são mapeados a partir dos relacionamentos entre as tabelas. Os mapeamentos gerados são registrados em um arquivo utilizando a linguagem *R2O* (BARRASA; CORCHO; GÓMEZ-PÉREZ, 2004). Algumas limitações da ferramenta são sua compatibilidade restrita aos Sistemas Gerenciadores de Bancos de Dados (SGBDs)*Oracle* e *MySQL*, a impossibilidade de integração de uma ontologia pré-existente ao mapeamento, e a ausência de interface para visualização e consulta dos dados ligados gerados.

O trabalho apresentado por SZEKELY, HEJJA e BUCHMANN (2011) descreve uma aplicação *Web* desenvolvida em linguagem*PHP*, utilizando a *API RAP* (RDF API for PHP), que efetua o mapeamento para *RDF* de uma BDR da área de recursos humanos de uma empresa específica e possibilita a visualização e consulta dinâmica dos dados ligados gerados. O aplicativo utiliza a plataforma *D2RQ*(BIZER; CYGANIAC, 2006) para o mapeamento e a

ISBN 978-85-61115-09-8

visualização e consulta de dados ligados. Além de ter escopo muito limitado, o aplicativo não permite a customização do mapeamento ou a incorporação de uma ontologia.

A geração automática de mapeamentos entre BDR e *RDF* utilizando as linguagens *R2RML* ou *Direct Mapping* é o propósito do *plugin* para ambiente *Eclipse* apresentado por SALAS et al. (2011). São oferecidos ao usuário diferentes algoritmos de mapeamento, os quais podem ser escolhidos por meio de uma interface gráfica, bem como outros algoritmos podem ser adicionados. A ferramenta, entretanto, não possibilita a customização do mapeamento ou a incorporação de uma ontologia a este, o que deve ser feito manualmente pelo usuário por meio da edição do arquivo de mapeamento gerado.

O *Triplify*(AUER et al., 2009) é um pequeno *plugin*, desenvolvido em linguagem *PHP*, com a finalidade de converter em tuplas os dados relacionais manipulados por aplicações e páginas da *Web* existentes, e publicá-los na forma de dados ligados (*linked data*). A conversão dos dados relacionais para *RDF* é efetuada mediante a reescrita, na forma de triplas (tabelas como classes e colunas como *URI*), dos resultados provenientes de consultas *SQL* efetuadas nas respectivas BDRs. Dessa forma, uma limitação da ferramenta é que ela não possui suporte a consultas em linguagem *SPARQL*.

O *framework Iconomy* (VAVLIAKIS et al., 2011)tem por objetivo fazer a transformação de BDRs em dados ligados por meio da criação automática de instâncias em uma ontologia *OWL* existente. Oferece um conjunto de interfaces gráficas, por meio das quais é possível: customizar e editar o mapeamento entre uma BDR e uma ontologia em formato *OWL* para criação das instâncias; a construção, de forma interativa, de consultas *SPARQL*; e a inserção de restrições e regras na ontologia, de forma a proporcionar suporte à inferência. As limitações da ferramenta ficam por conta da compatibilidade restrita aos SGBDs *Oracle* e *MySQL*, e do tipo de mapeamento gerado, que é estático (baseado em *ETL*). Além disso, a formulação de consultas *SPARQL* fica limitada às opções oferecidas pela interface gráfica.

O *LOD2 Stack* (AUER et al., 2012) é um *framework* que agrupa uma série de ferramentas para a publicação de dados ligados, abrangendo as tarefas de extração, consulta e exploração, criação, descoberta semi-automática de ligações (*links*) entre fontes de dados e, enriquecimento e reparação de bases de conhecimento. Uma característica relevante do *framework* é a adoção da plataforma *D2RQ*(BIZER; CYGANIAC, 2006) para a extração de dados ligados a partir de BDRs, bem como para a visualização e consultas. Como limitações, não é possível a incorporação de ontologia ao mapeamento entre BDR e *RDF*, e a aplicação está disponível somente para plataforma *Linux*. Além disso, a maior parte das ferramentas oferecidas no *framework* são aplicações disponíveis somente de forma *on-line* na *Web*.

O *Dartgrid*(WU et al., 2006) é um *framework* que oferece um conjunto de ferramentas com o propósito de integrar BDRs heterogêneas usando a *Web* Semântica. Uma das ferramentas oferecidas permite ao usuário, por meio de interface gráfica, construir mapeamentos entre oesquema de uma BDR e uma ontologia. Outra ferramenta possibilita, também por interface gráfica, gerar consultas *SPARQL* baseadas nos termos da ontologia mapeada. Os aspectos negativos destas ferramentas ficam por conta de sua interface gráfica, que está em idioma chinês, e da impossibilidade de edição manual do arquivo de mapeamento gerado. Além disso, o *framework* foi projetado para um domínio específico, que consiste na interconexão de várias BDRs referentes à Medicina Tradicional Chinesa.

Por fim, o *Ne-on Toolkit*(HAASE, et al., 2007) é um *framework* para a construção e o gerenciamento de ontologias que conta com uma grande quantidade de ferramentas e *plugins* para estas finalidades. Apesar de ter seu foco na construção de ontologias, o *framework* conta com um *plugin* específico, chamado de *ODE Mapster*(RODRIGUEZ; GÓMES-PÉREZ, 2006), que oferece uma interface gráfica para o mapeamento semi-automático entre ontologias e BDRs, expresso em linguagem *R2O* (BARRASA; CORCHO; GÓMEZ-PÉREZ,

ISBN 978-85-61115-09-8

2004). A aplicação é compatível somente com os SGBDs *Oracle* e *MySQL*, e não oferece interface para a publicação (geração e visualização e consulta) dos dados ligados mapeados.

O diferencial da abordagem proposta neste artigo, em comparação aos trabalhos relacionados, está na automatização, implementada pela interface gráfica, das ferramentas para a geração e customização do mapeamento entre BDR e *RDF*, e para a visualização e consulta dos dados ligados obtidos a partir desse mapeamento. A Tabela 1 apresenta um comparativo entre a abordagem proposta,chamada *RDB2LOD*, e os trabalhos relacionados.

TABELA 1–Comparativo entre a Abordagem RDB2LOD e Trabalhos Relacionados.

QUESITO	DB2OWL	Web App Recursos Humanos	Plugin Eclipse	Triplify	Iconomy	LOD2 Stack	DartGrgrid	Ne-On Toolkit	Abordagem RDB2LOD
Método de Mapeamento	Dinâmico/ Estático	Dinâmico	Dinâmico/ Estático	Dinâmico/ Estático	Estático	Dinâmico/ Estático	Dinâmico/ Estático	Dinâmico/ Estático	**Dinâmico/ Estático**
Permite mapeamentos genéricos?	Sim	Não	Sim	Sim	Sim	Sim	Não	Sim	**Sim**
Linguagem do arquivo de mapeamento	R2O	D2RQ	R2RML/ Direct Mapping	SQL Views	Específica da Aplicação	D2RQ	Específica da Aplicação	R2O	**D2RQ**
SGBDs Compatíveis	MySQL/ Oracle	MySQL/ Oracle	MySQL/ Oracle	Diversos	MySQL/ Oracle	MySQL/ Oracle	MySQL/ Oracle	MySQL/ Oracle	**MySQL/MS-SQL Server**
Ambiente de Execução	Não Informado	Web	Eclipse	Web	Java Runtime	Linux (Debian) / Web	Java Runtime	Java Runtime	**Java Runtime**
Arquivo de mapeamento editável?	Sim	Não	Sim	Sim	Sim	Sim	Não	Sim	**Sim**
Incorporação de ontologia pelo usuário?	Não	Não	Não	Não	Sim	Não	Não	Sim	**Sim**
Visualização e consulta dos dados gerados	Não	Sim	Não	Sim	Sim	Sim	Sim	Não	**Sim**

Fonte: Autor

Destaca-se também como diferenciais dessa abordagem, a compatibilidade com o SGBD *MS SQL Server*, e a ferramenta de implementação do método para incorporação, de forma semi-automática (também por meio de interface gráfica), de uma ontologia ao arquivo de mapeamento, esta compatível com as recomendações de padrão do *W3C*.

3Abordagem para a publicação de Dados Ligados obtidos a partir de BDRs

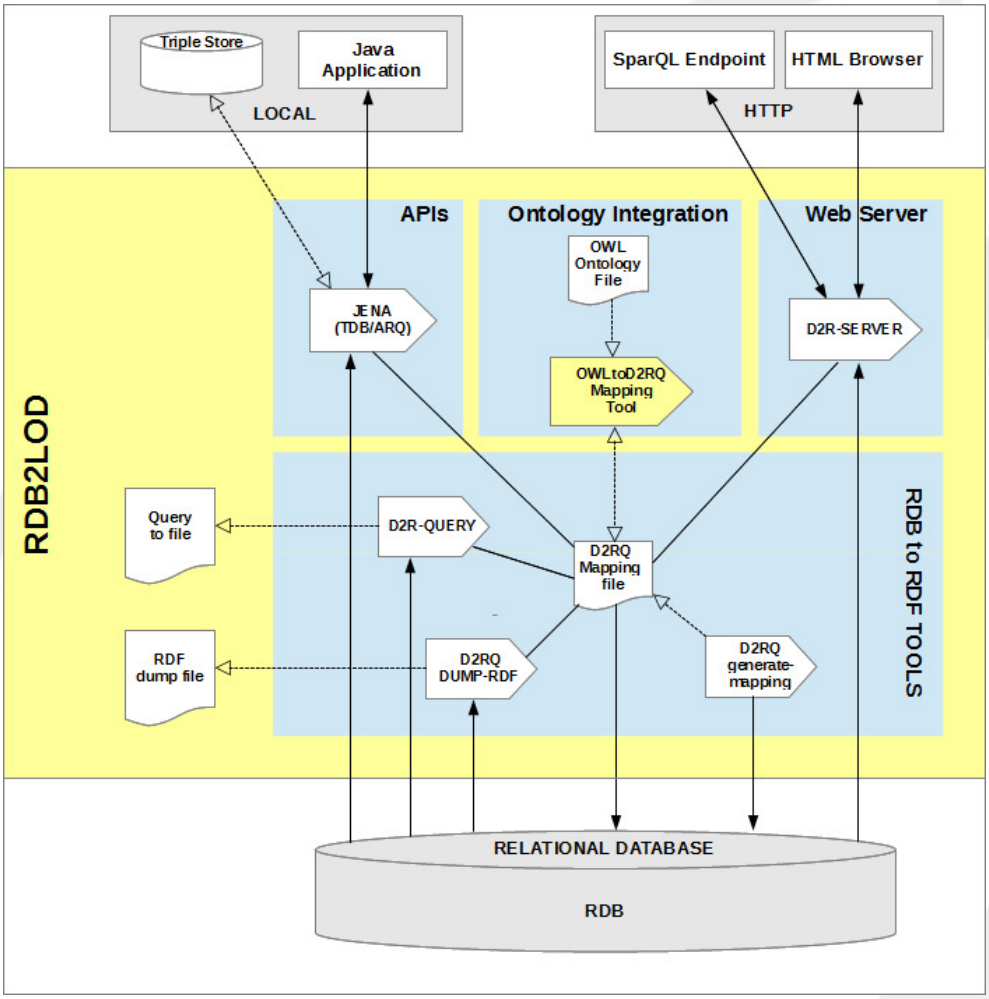
As ferramentas atualmente disponíveis para efetuar a conversão de dados mantidos em BDRs para dados ligados na forma de triplas *RDF* possuem diversas limitações, que exigem do usuário um elevado nível de conhecimento técnico e de interação manual. Uma dessas limitações é que várias destas ferramentas são configuradas e acionadas através de linha de comando do sistema operacional. Outra limitação é a necessidade de edição manual do arquivo de mapeamento entre BDR e *RDF* para a substituição do vocabulário padrão (nomes de tabelas como classes e nomes de colunas como propriedades) por elementos de uma ontologia de domínio fornecida pelo usuário.

A abordagem apresentada neste artigo, chamada de *RDB2LOD*, busca suprir essas limitações ao oferecer uma aplicação que integra alguma destas ferramentas, por meio de uma interface gráfica onde é possível configurá-las e acioná-las, e de uma nova ferramenta que possibilita a customização do arquivo de mapeamento para incorporação de uma ontologia do domínio fornecida pelo usuário. Dessa forma, ficam bastante reduzidas as necessidades de

ISBN 978-85-61115-09-8

interação manual e de conhecimento técnico por parte do usuário. A Figura 2 apresenta uma visão geral da abordagem *RDB2LOD*, a qual é composta de duas camadas.

Figura 2 – Visão geral da abordagem *RDB2LOD*



Fonte: Autor

A camada inferior da abordagem conta com uma ferramenta para efetuar a conexão com a BDR, conforme parâmetros fornecidos pelo usuário. Por meio desta ferramenta é obtida a estrutura da BDR e gerado um arquivo de mapeamento, a partir do qual os dados da BDR serão convertidos para o formato de triplas *RDF*.As triplas geradas podem então ser utilizadas por outras duas ferramentas, oferecidas também nesta camada, para a geração de *datasets*, ou então serem repassadas para as ferramentas da camada superior.

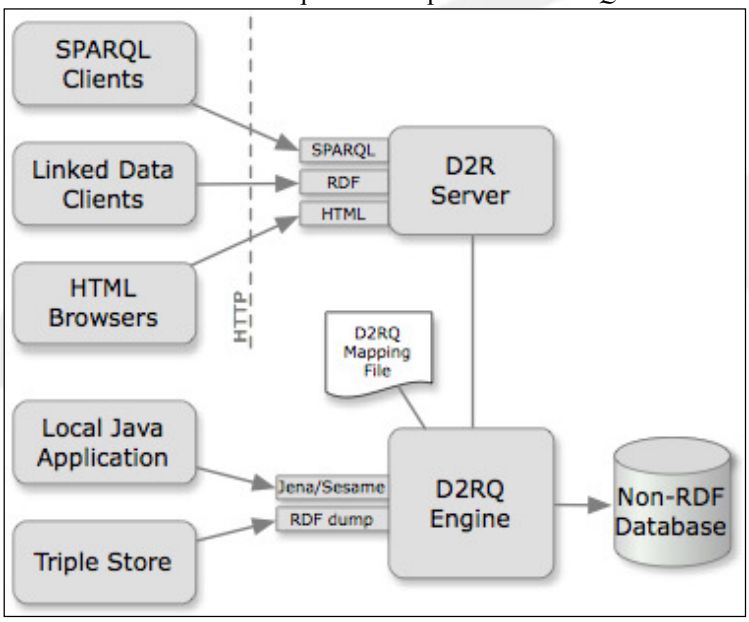
Para a seleção de tais ferramentas, dentre os critérios utilizados no processo de avaliação, destacam-se: a possibilidade de edição manual do arquivo de mapeamento gerado (para efetuar a customização e incorporação de ontologia); linguagem de mapeamento empregada pela ferramenta compatível com as linguagens indicadas como recomendação de padrão pelo *W3C* (*R2RML* e *Direct Mapping*); a possibilidade de efetuar mapeamentos tanto dinâmicos (*on-line*) quanto estáticos (baseados em *ETL*); a compatibilidade da ferramenta com os SGBDs *MySQL* e *MS SQL Server*; e, a ferramentaser distribuída na forma de licença pública (*open source/GPL*).

Assim, a plataforma *D2RQ* (BIZER; CYGANIAC, 2006) foi a escolhida, por atender a todos estes critérios. Ela oferece um conjunto de ferramentas para a geração de dados ligados, na forma de triplas *RDF*, a partir do acesso a BDRs. Seus principais componentes são a *D2RQ Mapping Language*, o *D2RQ Engine* e o *D2R Server*. Por meio dessas ferramentas é possível

ISBN 978-85-61115-09-8

acessar o conteúdo de BDRs na forma de dados ligados (*Linked Data*), e efetuar consultas *SPARQL* sobre estes dados, localmente ou pela *Web*. Também possibilita a geração de *datasets* em formato *RDF* a partir da descarga de todo o conteúdo (*dump*) da BDR acessada.Uma visão geral de sua arquitetura é apresentada na Figura 3.

FIGURA 3- Arquitetura da plataforma *D2RQ*.

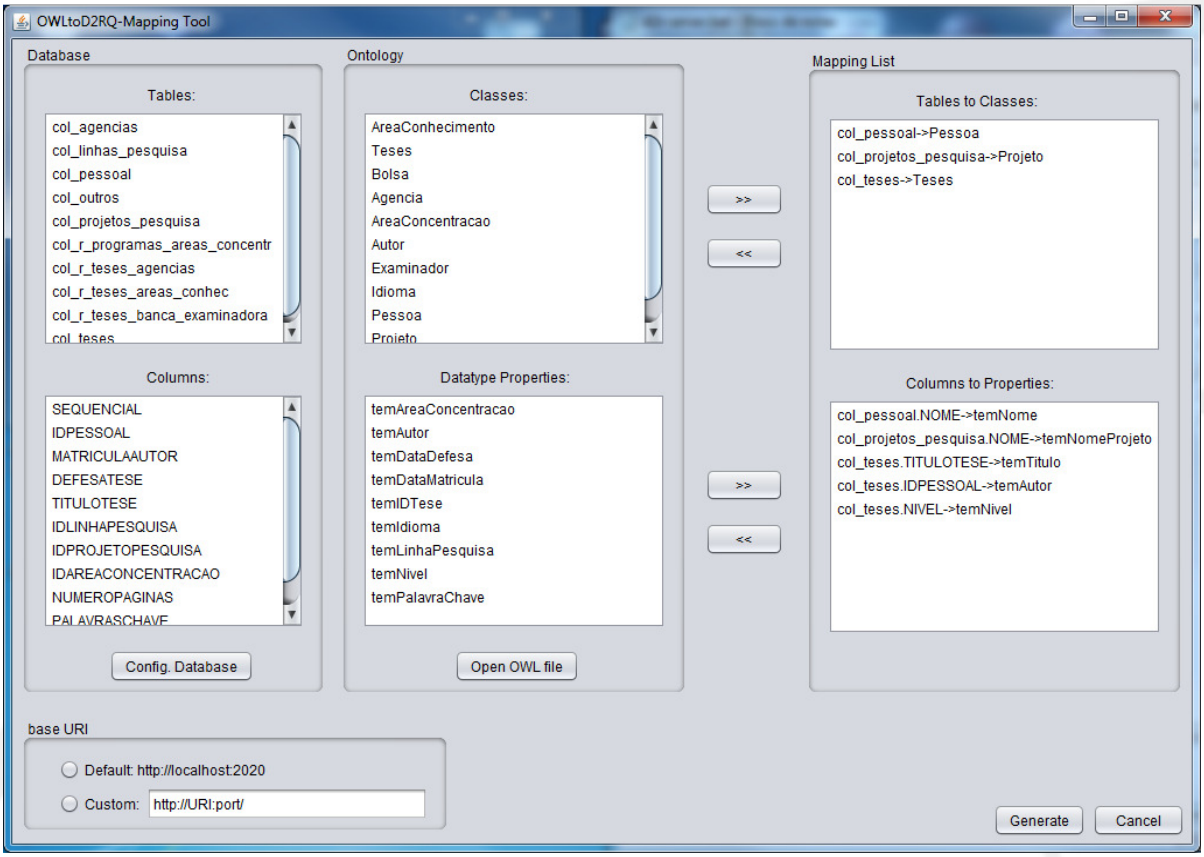


Fonte: BIZER; CYGANIAC, 2006.

Já a camada superior conta com três componentes distintos: a ferramenta *D2R-Server*, que possibilita a publicação do conteúdo de uma BDR na *Web* Semântica de forma dinâmica, ou seja, efetuando sob demanda a conversão dos dados da BDR mapeada para o formato de triplas *RDF*, sem a necessidade de replicar o conteúdo dessa base de dados em um armazenamento dedicado de dados ligados (*triple store*); as *APIs* do *framework Jena*, com o intuito de possibilitar o desenvolvimento de aplicações externas pelo usuário; e a ferramenta *OWLtoD2RQ-Mapping*, desenvolvida nesta abordagem, que tem por finalidade customizar, de forma interativa e semi-automática, o arquivo de mapeamento entre BDR e *RDF*de forma a permitir a incorporação de uma ontologia do domínio fornecida pelo usuário. A Figura 4 apresenta a interface gráfica desta ferramenta.

ISBN 978-85-61115-09-8

Figura 4 – Interface gráfica da ferramenta *OWLtoD2RQ-Mapping* desenvolvida na abordagem



Fonte: Autor

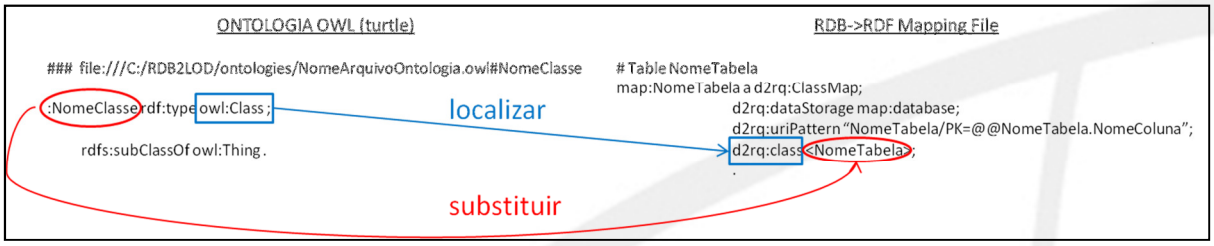
O métodopara efetuar a incorporação dos elementos de uma ontologia *RDFS* ao arquivo de mapeamento entre BDR e *RDF*, implementado pela ferramenta *OWLtoD2RQ-Mapping*, consiste basicamente da edição deste arquivo em três etapas.

Na primeira etapa são inseridas algumas linhas (fixas) no final do cabeçalho do arquivo de mapeamento (cláusulas *@prefix*), para a inclusão de parâmetros relacionados à customização a ser efetuada.

Na segunda etapa são substituídos, no corpo do arquivo de mapeamento, os rótulos das classes mapeadas (que, por padrão, são os nomes das respectivas tabelas do BDR) pelas classes da ontologia *RDFS* associadas, pelo usuário, por meio da interface gráfica da ferramenta. Para cada substituição a ser processada no arquivo de mapeamento, é efetuada a localização da cláusula "*d2rq:class*", acompanhada do nome da respectiva tabela, e, quando encontrada, esta é substituída pelo nome da classe da ontologia que foi associada a esta tabela.A Figura 5 permite uma representação deste procedimento.

ISBN 978-85-61115-09-8

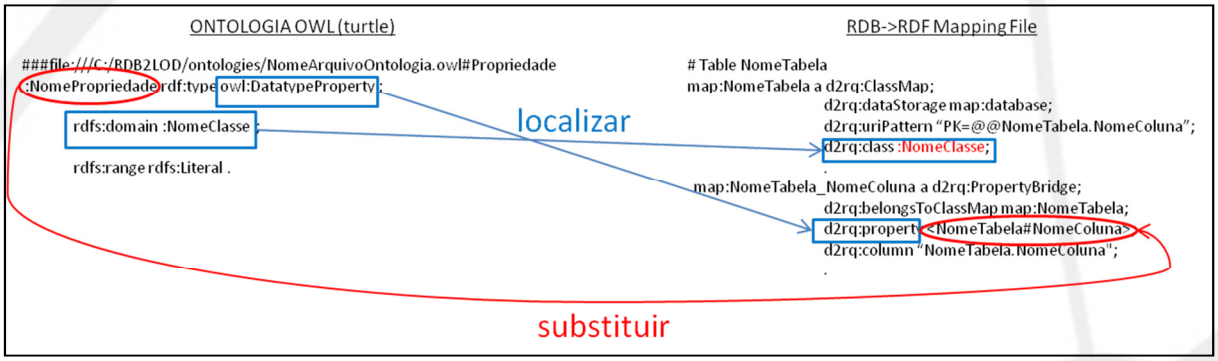
Figura 5 – Localização e substituição do nome da tabela pela classe da ontologia *RDFS* associada



Fonte: Autor

Na terceira etapa, são localizados e substituídos, para cada classe (tabela BDR) no arquivo de mapeamento, os rótulos das propriedades, ou seja, os nomes das colunas BDR mapeadas. Para cada propriedade da ontologia *RDFS* fornecida, são localizados no arquivo de mapeamento a correspondente classe (cláusula *d2rq:class*) e a respectiva coluna da tabela BDR associada a esta propriedade (cláusula *d2rq:property*) para, então, efetuar a substituição do nome da coluna BDR pelo nome da propriedade *RDFS* associada. A Figura 6 permite visualizar esta alteração.

Figura 6 – Localização e substituição do nome da coluna BDR pela propriedade *RDFS* associada



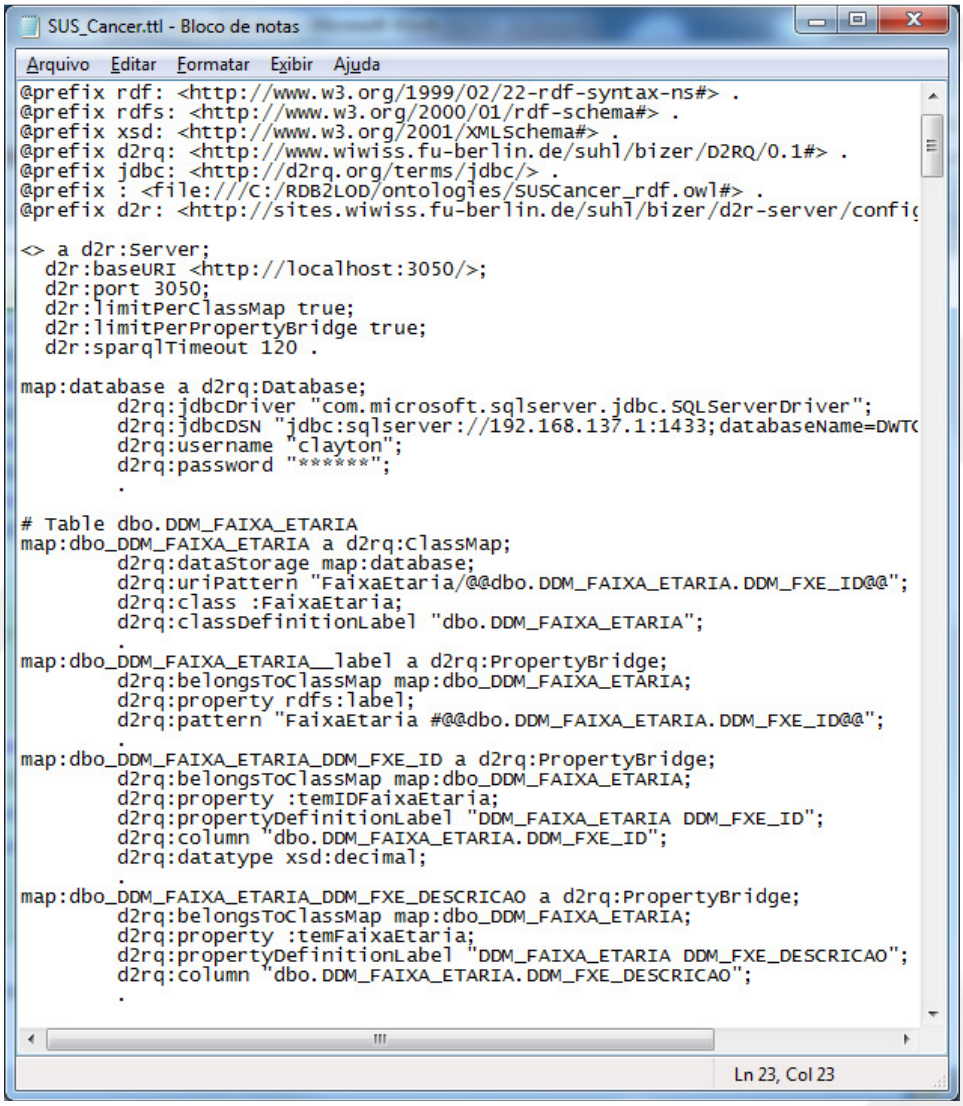
Fonte: Autor

Ao final dessas três etapas de alteração, o arquivo de mapeamento entre BDR e *RDF*, já com os elementos da ontologia *RDFS* incorporados, pode ser utilizado para gerar os dados ligados na forma de triplas *RDF*. Assim, o valor (ou objeto) de cada tripla *RDF* a ser gerada corresponderá a um campo de uma coluna (predicado) referente a uma tabela (sujeito) da BDR mapeada, ou ainda ao campo chave de outra tabela relacionada (expressa por meio de uma *URI*), quando esta coluna (predicado) for uma chave estrangeira da tabela mapeada. Este arquivo customizado de mapeamento pode ser visualizado na Figura 7.

A partir desse método será possível proporcionar um ganho de expressividade às triplas *RDF* geradas, ao apresentar significados bem definidos para os sujeitos, predicados e objetos destas, obtidos da customização do mapeamento da BDR com a incorporação da ontologia do respectivo domínio.

ISBN 978-85-61115-09-8

Figura 7 – Arquivo customizado de mapeamento de uma BDR para o modelo de dados *RDF*.



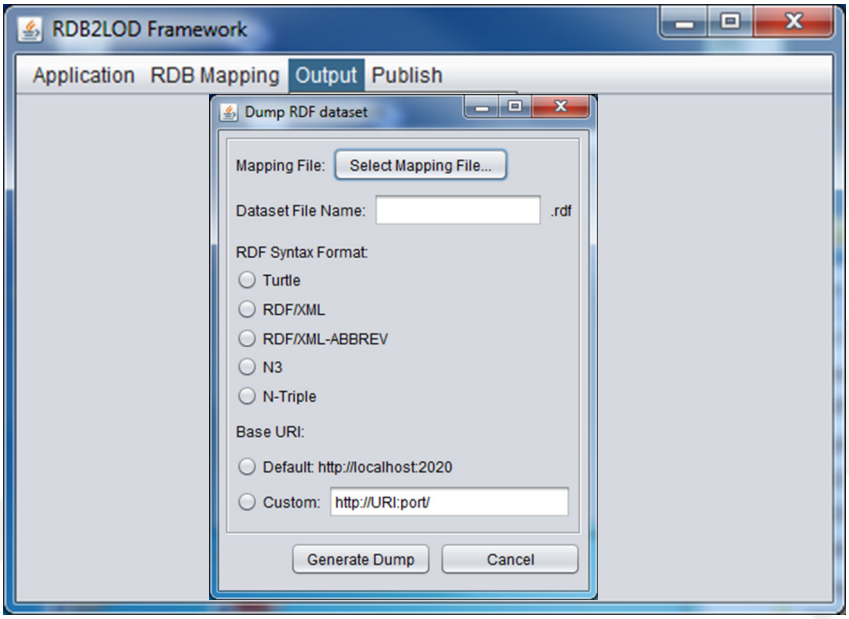
Fonte: Autor

Para prover as interfaces gráficas necessárias para a automatização e integração das ferramentas que compõem a abordagem, foi desenvolvido um aplicativo, em linguagem *Java*, chamado *RDB2LOD*. Este aplicativo possibilita a coleta de parâmetros, por meio de campos e listas ou botões de seleção, com a finalidade de automatizar a elaboração e execução das linhas de comando para acionamento e operação de cada uma das ferramentas aplicadas na abordagem.

Para cada uma das opções do menu principal do aplicativo, é aberta uma janela para a coleta de parâmetros que são armazenados em variáveis locais do código *Java*, de forma a permitir a construção e execução automatizadas da linha de comando da respectiva ferramenta. A Figura 8 mostra uma tela do aplicativo.

ISBN 978-85-61115-09-8

Figura 8 – Tela do aplicativo desenvolvido para integração das ferramentas aplicadas na abordagem.



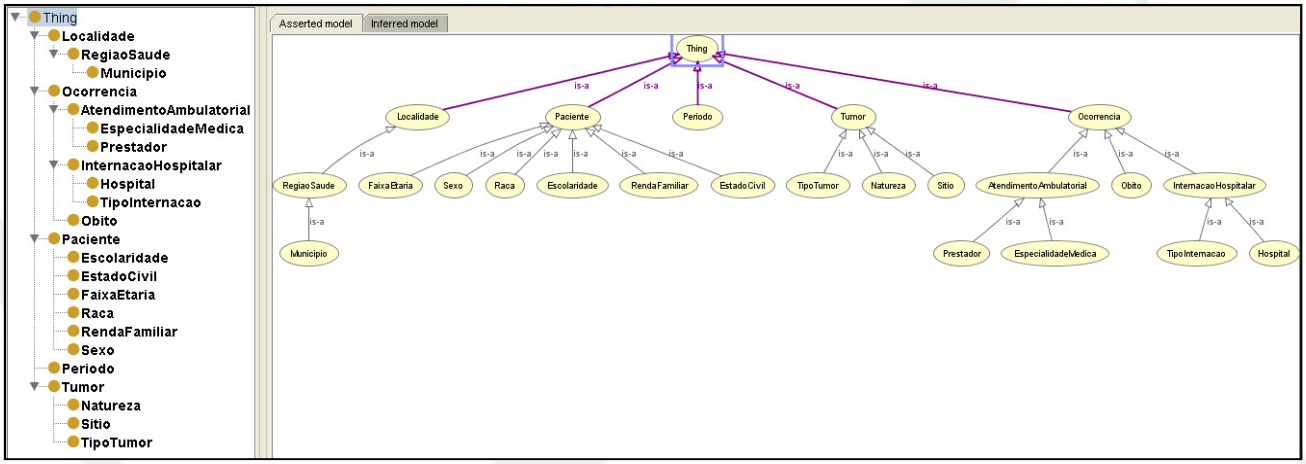
Fonte: Autor

4Estudo de Caso

A abordagem *RDB2LOD* foi aplicada a um estudo de caso em que foi utilizada uma BDR que contém dados referentes ao acompanhamento (mortes, internações hospitalares e atendimentos ambulatoriais) de portadores de câncer, a qual é mantida pelo Sistema Único de Saúde – SUS, e encontra-se na forma de um *datawarehouse* (chamado DW-SUS), onde são armazenados os dados históricos coletados de três sistemas de informações distintos: SIM (Sistema de Informações sobre Mortalidade); SIA (Sistema de Informações sobre Atendimento Ambulatorial); e SIH (Sistema de Informações Hospitalares).

Para a aplicação da abordagem, com o propósito de gerar e publicar os dados ligados obtidos a partir do mapeamento customizado dessa BDR, primeiramente foi obtida uma ontologia do domínio "Acompanhamento de Tumores Cerebrais", mostrada na Figura 9.

Figura 9 – Tela do aplicativo desenvolvido para integração das ferramentas aplicadas na abordagem



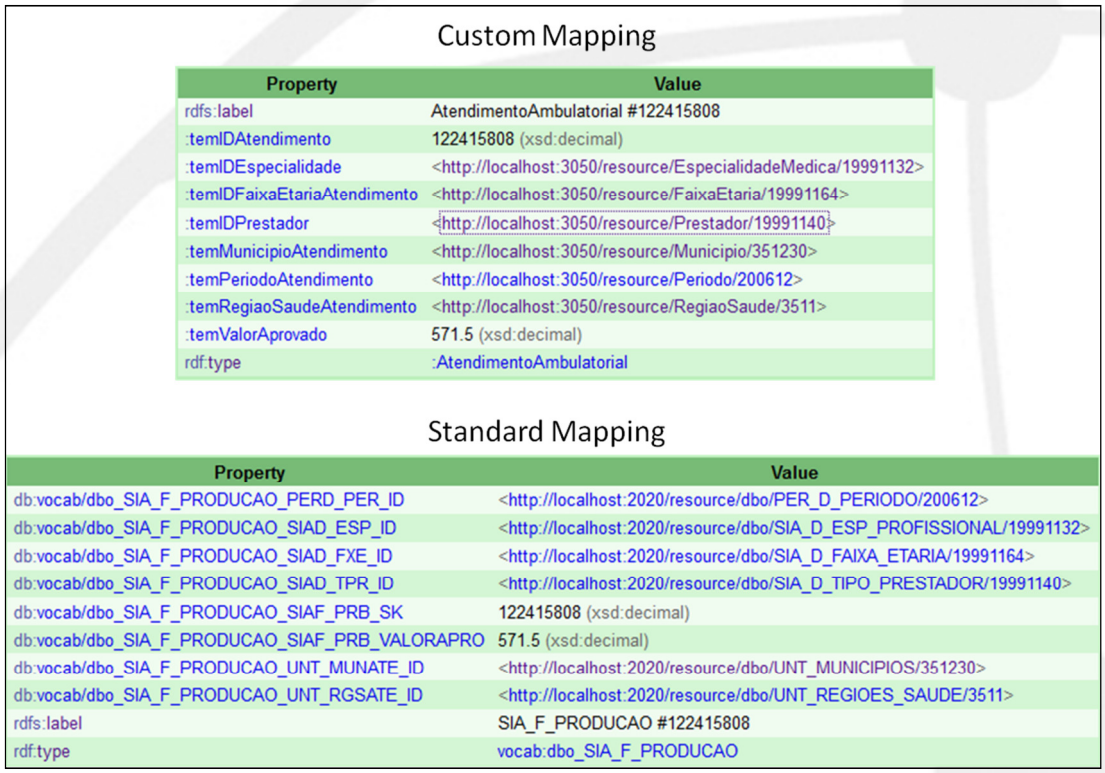
Fonte: Autor

ISBN 978-85-61115-09-8

A partir do arquivo dessa ontologia *RDFS* (em formato *OWL*), e dos parâmetros para conexão à BDR, foi possível efetuar a customização do arquivo de mapeamento, de forma semi-automática, por meio da ferramenta *OWLtoD2RQ-Mapping*, que efetuou a substituição do vocabulário padrão pelos termos da ontologia do domínio.

Em seguida, foram comparadas as visualizações dos dados ligados gerados, tanto a partir do mapeamento padrão da BDR, quanto a partir do mapeamento customizado (com a incorporação da ontologia do domínio). A Figura 10 apresenta uma visualização de dados referentes a uma internação hospitalar.

Figura 10 – Comparação entre visualizações de dados referentes a uma internação hospitalar, efetuadas com e sem a customização do arquivo de mapeamento.



Custom Mapping

Property	Value
rdfs:label	AtendimentoAmbulatorial #122415808
:temIDAtendimento	122415808 (xsd:decimal)
:temIDEspecialidade	<http://localhost:3050/resource/EspecialidadeMedica/19991132>
:temIDFaixaEtariaAtendimento	<http://localhost:3050/resource/FaixaEtaria/19991164>
:temIDPrestador	<http://localhost:3050/resource/Prestador/19991140>
:temMunicipioAtendimento	<http://localhost:3050/resource/Municipio/351230>
:temPeriodoAtendimento	<http://localhost:3050/resource/Periodo/200612>
:temRegiaoSaudeAtendimento	<http://localhost:3050/resource/RegiaoSaude/3511>
:temValorAprovado	571.5 (xsd:decimal)
rdf:type	:AtendimentoAmbulatorial

Standard Mapping

Property	Value
db:vocab/dbo_SIA_F_PRODUCAO_PERD_PER_ID	<http://localhost:2020/resource/dbo/PER_D_PERIODO/200612>
db:vocab/dbo_SIA_F_PRODUCAO_SIAD_ESP_ID	<http://localhost:2020/resource/dbo/SIA_D_ESP_PROFISSIONAL/19991132>
db:vocab/dbo_SIA_F_PRODUCAO_SIAD_FXE_ID	<http://localhost:2020/resource/dbo/SIA_D_FAIXA_ETARIA/19991164>
db:vocab/dbo_SIA_F_PRODUCAO_SIAD_TPR_ID	<http://localhost:2020/resource/dbo/SIA_D_TIPO_PRESTADOR/19991140>
db:vocab/dbo_SIA_F_PRODUCAO_SIAF_PRB_SK	122415808 (xsd:decimal)
db:vocab/dbo_SIA_F_PRODUCAO_SIAF_PRB_VALORAPRO	571.5 (xsd:decimal)
db:vocab/dbo_SIA_F_PRODUCAO_UNT_MUNATE_ID	<http://localhost:2020/resource/dbo/UNT_MUNICIPIOS/351230>
db:vocab/dbo_SIA_F_PRODUCAO_UNT_RGSATE_ID	<http://localhost:2020/resource/dbo/UNT_REGIOES_SAUDE/3511>
rdfs:label	SIA_F_PRODUCAO #122415808
rdf:type	vocab:dbo_SIA_F_PRODUCAO

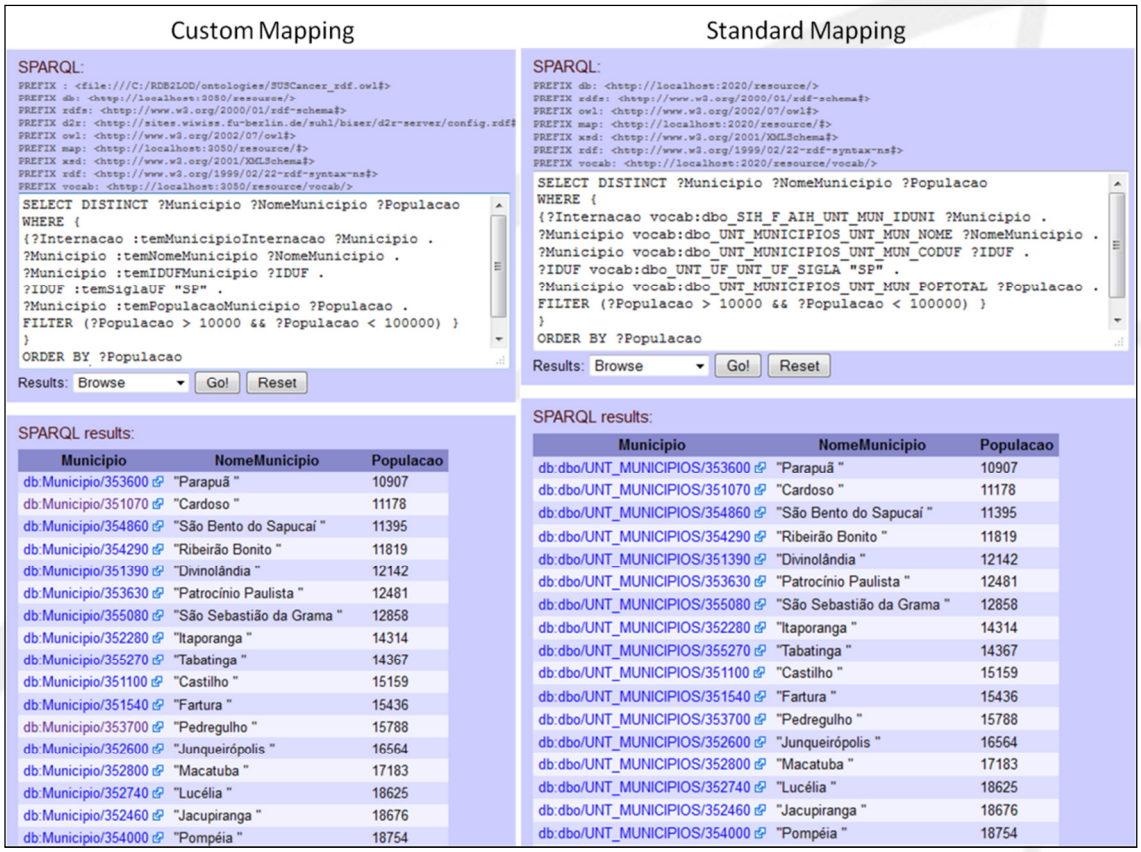
Fonte: Autor

Na Figura 10 é possível perceber que os rótulos das propriedades, quando gerados a partir do mapeamento customizado da BDR, apresentam um significado bem definido (aproximados à linguagem natural) para os respectivos valores apresentados, ao contrário dos rótulos de propriedades gerados a partir do mapeamento padrão, onde são utilizados nomes das colunas da BDR, que neste caso não impedem, mas dificultam a interpretação dos respectivos valores apresentados, pois dependem da compreensão de seus significados.

Por fim, foi efetuada uma comparação entre uma consulta *SPARQL* formulada a partir do mapeamento padrão, e a mesma consulta formulada a partir do mapeamento customizado da BDR. A Figura 11 mostra as consoles *SPARQL* com as consultasformuladas e os respectivos resultados obtidos, para ambos os mapeamentos. É importante observar as diferenças entre as sintaxes das consultas *SPARQL* formuladas, tanto para o mapeamento customizado, quanto para o mapeamento padrão. A utilização dos termos da ontologia do domínio no mapeamento customizado, em substituição aos nomes de colunas BDR no mapeamento padrão, permite uma melhor compreensão das propriedades (predicados das triplas *RDF*), que se aproximam da linguagem natural e facilitam uma eventual correção ou posterior edição dessa consulta.

ISBN 978-85-61115-09-8

Figura 11 – Comparação entre consultas *SPARQL*, efetuadas com e sem a customização do arquivo de mapeamento.



Fonte: Autor

A partir das comparações apresentadas, tanto na visualização quanto na consulta *SPARQL*, é possível afirmar que a aplicação da abordagem *RDB2LOD* proporcionou uma melhor expressividade às triplas *RDF* geradas, ao apresentar significados bem definidos (aproximados à linguagem natural) para os sujeitos, predicados e objetos destas, que foram obtidos da customização do mapeamento da *BDR* com a incorporação de uma ontologia do domínio.

A aplicação da abordagem *RDB2LOD* também proporcionou uma interação do usuário por meio de interfaces gráficas, eliminando a necessidade de interação manual para a configuração e operação das ferramentas aplicadas ao processo de publicação de dados ligados. Desta forma, a customização do mapeamento pelas associações entre tabelas e colunas da BDR e as classes e propriedades da ontologia, que até então era feita manualmente, passa a ser feita de forma automatizada por meio desta abordagem, o que reduz o tempo gasto e a necessidade de conhecimento técnico da linguagem do arquivo de mapeamento gerado pela ferramenta aplicada. A customização do mapeamento de bases de dados com um número elevado de tabelas e colunas, que seria impraticável pelo método convencional (manualmente), passa a ser possível por meio desta abordagem.

5Conclusão

Estaabordagem vem preencher uma lacuna encontrada nas ferramentas disponíveis para a geração e publicação de dados ligados a partir de dados estruturados persistidos em bases de dados relacionais, que é a falta de uma interface gráfica que automatize todo o processo e integre as diferentes ferramentas aplicadas, bem como gerar as triplas *RDF* com

ISBN 978-85-61115-09-8

base em uma ontologia de domínio, de forma a conferir maior expressividade às visualizações e consultas desses dados.

A abordagem *RDB2LOD* possibilita a integração de ferramentas e a automatização do processo de publicação de dados ligados obtidos a partir de BDRs, ao oferecer uma interface gráfica que reduz a necessidade de conhecimento técnico e elimina a necessidade de interação manual do usuário.

A partir do método e da ferramenta gráfica para customização do mapeamento entre a BDR e o formato *RDF*, com a incorporação de uma ontologia do domínio foi possível proporcionar melhor expressividade às triplas *RDF* geradas, que passaram a apresentar significados bem definidos (aproximados à linguagem natural) para os sujeitos, predicados e objetos. Isso impactou de forma positiva a visualização dos dados ligados publicados, a geração de *datasets*, e a elaboração de consultas *SPARQL*. Da mesma forma, a aplicação de uma ontologia do domínio nesta customização pode evitar a geração de triplas irrelevantes ou desnecessárias, possibilitando melhor visualização e exploração dos dados ligados obtidos.

Como trabalhos futuros, que possam complementar ou expandir a abordagem RDB2LOD, são sugeridos: dotar a ferramenta *OWLtoD2RQ-Mapping* de capacidade para gerar e customizar o mapeamento a partir de múltiplas BDRs, bem como ampliar a compatibilidade com outros SGBDs; permitir a associação entre tabelas da BDR e *Object Properties* da ontologia em formato *OWL* a ser incorporada na customização do mapeamento, de forma a estabelecer outras relações que não estavam implícitas no esquema desta BDR; e, possibilitar o mapeamento de regras, restrições e cardinalidade para as classes, durante a customização do mapeamento da BDR.

Para viabilizar a distribuição do aplicativo *RDB2LOD*na forma de licença pública (*open source/GPL*), recomenda-se ainda como trabalho futuro a correção e otimização de seu código-fonte (escrito em linguagem *Java*). O protótipo de aplicativo desenvolvido para validação da abordagem foi customizado para execução em ambiente *MS-Windows*. A execução em ambiente *Linux* pode requerer nova customização do aplicativo.

Por fim conclui-se que a abordagem *RDB2LOD* permitirá a publicação na *Web*, em grande escala, de dados ligados obtidos a partir de bases de dados relacionais, de forma a possibilitar amplo acesso a muitos usuários, o que hoje ocorre ainda de forma muito tímida por meio de algumas iniciativas.

Referências

ALLEMANG, D.; HENDLER, J. A. **Semantic web for the working ontologist:** modeling in RDF, RDFS and OWL. Burlington: Morgan Kaufmann, 2008.

ANTONIOU, G.; HARMELEN, F. **A semantic web primer.** 2ed. Cambridge: MIT Press, 2008.

AUER, S. et al. Triplify- light-weight linked data publication from relational databases.In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 18., 2009, Madrid. **Proceedings…** New York: ACM, 2009. p. 621-630.Disponível em: <http://doi.acm.org/10.1145/1526709.1526793>. Acesso em: 20 set. 2012.

AUER, S. et al. Managing the life-cycle of linked data with the LOD2 stack.In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 11., 2012, Boston. **Proceedings...** Berlin: Springer, 2012. p. 1-16. Disponível em: <http://lod2.eu/ BlogPost/1214-paper-about-lod2-stack-accepted-for-iswc.html>. Acesso em: 20 set. 2012.

ISBN 978-85-61115-09-8

BARRASA, J.; CORCHO, Ó.; GÓMEZ-PÉREZ, A. R2O, an extensible and semantically based database-to-ontology mapping language.In: WORKSHOP ON SEMANTIC WEB AND DATABASES, 2., 2004, Toronto. **Proceedings...**New York: Springer, 2004. p. 1069-1070.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, New York, v. 284, p. 28-37, Mai. 2001.

BIZER, C.; CYGANIAC, R. **D2RQ platform:** publishing relational databases on the semantic web. 2006.Poster presented at the 5th. Conference International Semantic Web, 2006, Athens. Disponível em:<http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/Bizer-Cyganiak-D2R-Server-ISWC2006.pdf>. Acesso em: 01 out. 2012.

CHOI, M.-Y. et al. Interoperability between a relational data model and an RDF data model.In:INTERNATIONAL CONFERENCE ON NETWORKED COMPUTING AND ADVANCED INFORMATION MANAGEMENT, 6., 2010, Seoul. **Proceedings...** Piscataway: IEEE, 2010. p. 335-340.

CULLOT, N.; GHAWI, R.; Y'ETONGNON, K. DB2OWL: a tool for automatic database-to-ontology mapping. In: ITALIAN SYMPOSIUM ON ADVANCED DATABASE SYSTEMS, 15., 2007, Torre Canne. **Proceedings…**Torre Canne: Dipartimento di Informatica, Università degli Studi Di Bari, 2007. p. 491-494. Disponível em: <http://dblp.unitrier.de/rec/bibtex/conf/sebd/CullotGY07>. Acesso em: 20 set. 2012.

HAASE, P. et al. **Ontology engineering and plugin development with the neon toolkit.**2007.Tutorial presented at the 6th International Semantic Web Conference, 2007, Busan.

HERT, M. et al. UpLink: a linked data editor for BDR-to-RDF data.In: INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 7., 2011, New York. **Proceedings...** New York: ACM, 2011. p. 159-162. Disponível em:<http://doi.acm.org/10.1145/2063518.2063539>. Acesso em: 20 set. 2012.

LING, H.; ZHOU, S. Translating relational databases into RDF.In:INTERNATIONAL CONFERENCE ON ENVIRONMENTAL SCIENCE AND INFORMATION APPLICATION TECHNOLOGY, 2., 2010, Wuhan. **Proceedings...** Piscataway: IEEE, 2010. p. 464-467.

LV, Y.; MA, Z. M. Transformation of relational model to RDF model.In: IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS, 2008, Singapore. **Proceedings...** Piscataway: IEEE, 2008. p. 506-511.

MACHADO, A. L.; PARENTE DE OLIVEIRA, J. M. DIGO: an open data architecture for e-government.In: ENTERPRISE DISTRIBUTED OBJECT COMPUTING CONFERENCE WORKSHOPS, 15., 2011, Helsinki. **Proceedings…** Piscataway: IEEE, 2011. p. 448-456.

MOHAMED, H.; JINCAI, Y.; QIAN, J. Towards integration rules of mapping from relational databases to semantic web ontology.In:INTERNATIONAL CONFERENCE ON WEB INFORMATION SYSTEMS AND MINING, 2010, Sanya. **Proceedings...** Piscataway: IEEE, 2010. p. 335-339.

RODRIGUEZ, J. B.; GÓMEZ-PÉREZ, A. Upgrading relational legacy data to the semantic web.INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 15., 2006, Edinburgh. **Proceedings...**New York: ACM, 2006. p. 1069-1070. Disponível em: <http://doi.acm.org/10.1145/1135777.1136019>. Acesso em: 20 set. 2012.

ISBN 978-85-61115-09-8

SAHOO, S. S. et al. **A survey of current approaches for mapping of relational databases to RDF.** [S.l.]: W3C, 2009. Disponível em:<http://www.w3.org/2005/Incubator/BDR2rdf/BDR2RDF_SurveyReport.pdf>. Acesso em: 20 set. 2012.

SALAS, P. E. et al. BDR2RDF plugin: relational databases to RDF plugin for eclipse.In:WORKSHOP ON DEVELOPING TOOLS AS PLUG-INS, 1., 2011, Honolulu. **Proceedings...**New York: ACM, 2011. p. 28-31. Disponível em: <http://doi.acm.org/10.1145/1984708.1984717>. Acesso em: 20 set. 2012.

SEQUEDA, J. F.; ARENAS, M.; MIRANKER, D. P. On directly mapping relational databases to RDF and OWL.In:INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 21., 2012, Lyon. **Proceedings...** New York: ACM, 2012. Disponível em:<http://doi.acm.org/10.1145/2187836.2187924>. Acesso em: 20 set. 2012.

SZEKELY, A.; HEJJA, A.; BUCHMANN, R. A. Mapping a relational database into a RDF repository.In:INTERNATIONAL SYMPOSIUM ON SYMBOLIC AND NUMERIC ALGORITHMS FOR SCIENTIFIC COMPUTING, 13., 2011, Timisoara. **Proceedings...**Piscataway: IEEE, 2011. p.175-182.

TURBAN, E. et al. **Administração de tecnologia da informação.** Rio de Janeiro: Campus, 2005.

VAVLIAKIS, K. N. et al. An integrated framework for enhancing the semantic transformation, editing and consultaing of relational databases.**Expert Systems with Applications**, Amsterdam, v. 38, n. 4,. p. 3844-3856,apr. 2011.

WU, Z. et al. Dartgrid: a semantic web toolkit for integrating heterogeneous relational databases.In: SEMANTIC WEB CHALLENGE AT 4TH INTERNATIONAL SEMANTIC WEB CONFERENCE, 4th., 2006, Athens. **Proceedings...** [S.l.]: Semantic Web Science Association, 2006. p. 750-763. Disponível em: <http://www.aifb.uni-karlsruhe.de/WBS/ywa/publications/wu06TCM_ISWC06.pdf>. Acesso em: 6 out. 2012.

ZHOU, S. Exposing relational database as RDF.In: INTERNATIONAL CONFERENCE ON INDUSTRIAL AND INFORMATION SYSTEMS, 2., 2010, Dalian. **Proceedings...** Piscataway: IEEE, 2010. p. 237-240.

ISBN 978-85-61115-09-8

Avaliação do ensino superior público no Brasil: protótipo de aplicação *linked data*

Rafael de Moura Speroni
rafaelsperoni@ifc-araquari.edu.br

Alexandre Moraes Ramos
alexandre.m.r@ufsc.br

Fernando A. Ostuni Gauthier
gauthier@egc.ufsc.br

Rafael Ramos da Luz
rafael.luz@cgu.gov.br

Claudelino Martins Dias Júnior
claudelino.junior@ufsc.br

Resumo
A fim de ampliar e estimular a participação e controle social sobre os processos de regulação, avaliação e supervisão da educação superior no sistema federal de educação, este trabalho apresenta uma potencial aplicação que permite o cidadão, de forma individual ou por meio da sociedade civil organizada, cruzar os dados dos portais do MEC, do INEP e das IFES para aferir corpo docente, projeto pedagógico, matriz curricular, resultados ENADE, informações gerais, de forma integrada e automatizada. A partir dos princípios de Web Semântica,*Linked Data* e do Modelo de Informações, descrito por Ramos e Marinho (2012), foi desenvolvida uma aplicação *Web mashup* para a visualização dos dados com auxílio de consultas SPARQL. Como resultado, este trabalho demonstra o potencial de aplicações desta natureza e conclui que para a efetiva participação dos cidadãos no sistema de nacional de avaliação do ensino superior é necessário que não só o Estado, mas principalmente as IFES disponibilizem os dados em formato aberto para o desenvolvimento de novas aplicações e o consequente aprimoramentos dos mecanismos de governança.

Palavras-chave: Avaliação do Ensino Superior. Linked Open Data. Mashup.

Abstract
In order to expand and encourage participation and public oversight over the process of regulation, evaluation and supervision of higher education in the federal educational system, this paper presents a potential application that allows citizens, individually or through organized civil society, to cross-reference the data from MEC, INEP, and IFES portals in order to assess in an integrated and automated waythe faculty staff,educational projects, syllabus content, ENADE'soutcomes, and other general information.Based on the principles of Semantic Web, Linked Data, and Information Model, described by Ramos and Marino (2012), we developed a Web mashup application for the visualization of the data with the aid of SPARQL queries. As a result, this work demonstrates the potential of such applications and concludes that for the effective citizen participation in the national evaluation system of higher education it is necessary that not only the state, but mostly Federal Institutions of Higher Education (IFES)make data available in an open format that allows the development of new applications and the consequent improvement of governance mechanisms.

Keywords: Higher Education Evaluation. Linked Open Data. Mashup.

ISBN 978-85-61115-09-8

Introdução

O presente trabalho decorre de estudos e pesquisas voltadas para obter uma maior compreensão e aprimoramento da aplicação da Lei de Acesso à Informação (LAI), nas Universidades Públicas Federais, a fim de ampliar a cultura de transparência e o nível de governança pública na Educação Superior do Brasil, desenvolvidos pelo Instituto de Pesquisas e Estudos em Administração Universitária (INPEAU)[18] da Universidade Federal de Santa Catarina (UFSC).

Na busca para se criar um modelo de informação, que servisse de referência para as Universidades Públicas Federais, no atendimento a divulgação de rotina, definida na Lei de Acesso à Informação – LAI (Lei nº 12.527/2011), e desta forma viabilizar a participação do cidadão, da comunidade acadêmica e da sociedade nos processos de regulação, avaliação e supervisão da educação superior brasileira, Ramos e Marinho (2012) destacam que a legislação vigente impôs um conjunto de preceitos/requisitos em termos legais que impactam diretamente o modelo atual de governança, disponibilização, acesso e uso dos dados da avaliação da educação superior no Brasil.

Ao se fazer uma revisão histórica, é possível notar que a educação superior de uma maneira geral está inserida em um ambiente complexo e com muitos desafios a fim de atender aos anseios da sociedade. Os desafios estão na garantia da qualidade do ensino, pesquisa e extensão, além de maior produtividade, eficiência, inovação e transparência na gestão universitária. Neste sentido, Schlickmann, Melo e Alperstedt (2008) destacam a importância do tema avaliação na busca da excelência da educação superior brasileira. Para Souza (2009), só um adequado relacionamento entre a universidade e a sociedade permitirá uma política de educação superior, de qualidade e perfeitamente ajustada às necessidades sociais.

Assim sendo, para se alcançar as metas, traçadas pelo Plano Nacional de Educação 2011-2020 do MEC (MEC, 2012), faz-se necessário aprimorar os mecanismos de governança, que garantam a participação e o controle social no Sistema Nacional de Avaliação da Educação Superior (SINAES).

Participar e controlar pressupõe que os cidadãos possam se envolver, discutir, intervir, propor e compartilhar ideias. A participação contínua da sociedade na gestão pública é um direito assegurado pela Constituição Federal de 1988, pela Lei de Acesso à Informação(Lei nº 12.527/2011), pela Política Nacional de Participação Social - PNPS (Decreto nº 8.243/ 2014) e, no caso específico da Educação Superior, pelo SINAES (Lei Nº. 10.861/2004), permitindo que os cidadãos não só participem da formulação das políticas públicas, mas, também, fiscalizem de forma permanente as ações do governo.

Para tanto, é essencial que o Estado promova e consolide a adoção de mecanismos que viabilizem a participação e o controle social em tais processos. A efetividade destes mecanismos de governança depende, dentre outros fatores, além do amparo legal e da disponibilização de dados abertos para cidadão, da estruturação dos conteúdos e organização da informação, especialmente a representação descritiva e temática para que possam ser criados aplicativos que integrem e gerem valor aos dados, viabilizando de fato a participação e o controle social no âmbito do SINAES.

Neste sentido, este artigo tem como proposta apresentar um protótipo de aplicação que permite a interligação de dados, de diferentes fontes de informação do SINAES[19]e das Instituições Federais de Ensino Superior(IFES), a fim de que o cidadão possa cruzaràs

[18] Projeto financiado pela Fapesc UNIVERSAL 2012. Termo de OUTORGA N. 20012-00009

[19]Disponibilizados a partir do portal do Ministério da Educação (MEC) e do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

ISBN 978-85-61115-09-8

informações dos processos de regulação, avaliação e supervisão da educação superior no sistema federal de educação. Para tanto, na primeira parte deste trabalho, são tratados os conceitos Dados Abertos, *Web* Semântica e outros a ele relacionados (*Linked Data*, RDF e *Web* de Dados), com o objetivo de compreender as tecnologias e os procedimentos necessários para a implantação de tal aplicação. Em seguida, é descrito o contexto da aplicação, com seus princípios norteadores. Por fim, é apresentada uma aplicação do tipo *mashup Web* para visualização de dados de avaliações de cursos superiores, onde dados provenientes de diferentes fontes são interligados e publicados segundo os princípios de *Linked Data*.

1 Dados Abertos

Segundo a definição da *Open Knowledge*[20], dados são abertos quando qualquer pessoa pode livremente usá-los, reutilizá-los e redistribuí-los, estando sujeito, no máximo, à exigência de creditar a sua autoria e compartilhá-los segundo a mesma licença. Devido às suas características, diversas iniciativas incentivam a produção e disponibilização de dados abertos, especialmente no âmbito governamental, como forma de aumentar a transparência e a participação dos cidadãos.

Os princípios básicos da abertura de dados são (OPEN KNOWLEDGE FOUNDATION, 2012): i) Disponibilidade e acesso: os dados devem estar disponíveis como um todo, preferencialmente via *download* na *internet*; ii) Reuso e redistribuição: os dados devem ser fornecidos sob termos que permitam reuso e redistribuição, incluindo a interligação com outros conjuntos de dados e suporte a processamento por máquinas e iii) Participação universal: qualquer pessoa deve ser capaz de usar, reusar e redistribuir os dados.

A adesão à publicação de dados abertos tem sido uma preocupação dos governos de vários países, e este comportamento é justificado porque os dados abertos viabilizam: Transparência e controle democrático; Participação; Auto-empoderamento; Produtos e serviços novos ou melhorados; Inovação; Melhora na eficiência dos serviços governamentais; Melhora na efetividade dos serviços governamentais; Mensuração de impacto das políticas; e Novos conhecimentos a partir de fontes de dados combinadas e padrões em grandes volumes de dados (OPEN KNOWLEDGE FOUNDATION, 2012).

A grande quantidade de dados abertos disponibilizadosna *Web* fez com que se desenvolvessem tecnologias no sentido de sua padronização e interligação. Desta forma, além de documentos interligados, a *Web* passa a contar com conjuntos de dados codificados em formatos abertos e padronizados, tais como CSV, JSON, XML e RDF (OPEN KNOWLEDGE FOUNDATION, 2012).

2 *Web* Semântica e *Linked Open Data*

A concepção da *Web* data do final da década de 1980, quando Tim Berners-Lee publicou uma proposta de um sistema para compartilhamento de documentos de pesquisa, chamando-a de *World Wide Web*(BERNERS-LEE, 1989). Assim, a*Web* é um espaço de informação na qual os itens de interesse, chamados recursos, são identificados por identificadores globais, os *Uniform Resource Ideintifiers* (URI). A Identificação, juntamente com a Interação e os Formatos, constituem as três bases arquiteturais da *Web* (JACOBS; WALSH, 2004).

A *Web* Semântica é uma extensão da *Web* tradicional, na qual as informações recebem significado bem definido, possibilitando que pessoas e computadores trabalhem em

[20]http://opendefiniton.org.

ISBN 978-85-61115-09-8

cooperação(BERNERS-LEE; HENDLER; LASSILA, 2001). Para seu funcionamento, os computadores devem ter acesso a coleções de informações e conjuntos de regras de inferência que possam usar para seu raciocínio automatizado. Duas tecnologias importantes para o desenvolvimento da *Web* Semântica são o *eXtensible Markup Language* (XML) e o *Resource Description Framework* (RDF).

O objetivo primordial da *Web* Semântica é a construção de um espaço global de dados baseado em padrões abertos, a *Web* de Dados, cuja proposta é ser uma extensão da *Web* atual, com base em um paradigma de publicação, não apenas documentos, mas também de dados (HEATH; BIZER, 2011).Para isso, é necessário que existam dados em grande quantidade disponíveis na Web em um formato padrão, acessíveis, interligados e gerenciáveis por ferramentas de *Web* Semântica. Esta coleção de conjuntos de dados inter-relacionados na Web pode também ser chamada de Dados Ligados (*Linked Data*).

O termo *Linked Data*, proposto por Berners-Lee (2006), refere-se a um estilo de publicar e interligar dados estruturados de diferentes fontes na *Web*. O assunto vem despertando considerável interesse acadêmico (BIZER; HEATH; BERNERS-LEE, 2009; BRADLEY, 2009; HEATH, 2011) e diversas ações estão sendo desenvolvidas no sentido da criação de repositórios de dados ligados.

Para que seja possível atingir o objetivo da *Linked Data*, de forma que os dados passem a fazer parte de um repositório global, Berners-Lee (2006) apresenta quatro regras, conhecidas como os princípios de*Linked Data*:

1) Deve-se usar URIs (*Uniform Resource Identifiers*) como nomes para as coisas.
2) Utilizar HTTP URIs (*Hyper Text Transfer Protocol*) de modo que as pessoas possam procurar por esses nomes.
3) Quando alguém procurar um URI, fornecer informações úteis usando os padrões RDF (*Resource Description Framework*) e SPARQL (linguagem de consulta).
4) Incluir links para outros URIs, a fim de que se possa descobrir mais coisas.

As quatro regras fornecem, portanto, uma receita básica para a publicação e conexão de dados usando a infraestrutura da web em aderência à sua arquitetura e padrões. Diferentemente da web de hipertextos, entretanto, onde as relações são ligações em documentos escritos em HTML, na web de dados as ligações são estabelecidas entre "coisas" arbitrárias descritas por RDF e as URIs identificam qualquer tipo de objeto ou conceito (BERNERS-LEE, 2006).

A *Web* de dados ligados pode ser vista como uma camada adicional que é estreitamente entrelaçada com a *Web* clássica de documentos e tem muitas propriedades em comum (BIZER, 2009):

a) Qualquer um pode publicar dados na *Web* de dados ligados.
b) *Links* conectam entidades, criando um grafo de dados global que se estende e possibilita a descoberta de novas fontes de dados.
c) Os dados são auto-descritos. Se uma aplicação encontra dados representados, usando um vocabulário que não lhe é familiar, pode resolver as URIs que identificam os termos do vocabulário para encontrar suas definições.
d) A *Web* de Dados é aberta, significando que aplicações podem descobrir novas fontes de dados seguindo os *links* durante sua execução.

As práticas de *Linked Data* oferecem, portanto, um mecanismo simples para combinação de múltiplas fontes na *Web* (HYLAND e WOOD, 2010), fornecendo uma gama de padrões internacionais e melhores práticas para a publicação, divulgação e reutilização de dados estruturados. Assim como muitos dados abertos não são dados ligados, os dados

ISBN 978-85-61115-09-8

publicados na forma de *Linked Data* não precisam, necessariamente, ser abertos. O termo *Linked Open Data* (LOD), segundo Berners-Lee (2006), refere-se a dados ligados (*Linked Data*) que são disponibilizados sob uma licença aberta, que não impede o seu livre reuso.

No ano de 2010, Tim Berners-Lee acrescentou ao documento de *Design Issues* (BERNERS-LEE, 2006), um sistema de avaliação dos dados com estrelas. O objetivo, segundo o autor, era encorajar as pessoas, especialmente os responsáveis por dados governamentais, ao caminho do bom *Linked Data*. Pelo sistema de estrela, detalhado no Quadro 1, os dados são classificados segundo seu poder e facilidade de uso para as pessoas.

Quadro 1 – As cinco estrelas dos dados abertos

★	Disponível na *web* (independente do formato), mas com uma licença aberta, para ser *Open Data*
★★	Disponível como dado estruturado legível por máquinas (ex. excel ao invés de uma imagem digitalizada de uma tabela)
★★★	Mesmo que (2), mas com formato não proprietário (ex. CSV ao invés de excel)
★★★★	Todas as anteriores, mais o uso de padrões abertos da W3C (RDF e SPARQL) para identificar as coisas, de maneira que as pessoas possam apontar para elas
★★★★★	Todas as anteriores, mais: Ligue seus dados aos dados de outras pessoas para prover contexto

Fonte: adaptado deBERNERS-LEE, 2006

Dentre as iniciativas relacionadas ao desenvolvimento dos dados ligados abertos, uma das mais importantes é o projeto *Linking Open Data*, comunidade fundada em 2007,cujo objetivo contínuo é promover à *Web* de Dados, por meio da identificação e publicação dos conjuntos de dados existentes, disponíveis sob licenças abertas, convertendo-os em RDF, de acordo com os princípios de *Linked Data*, e publicando-os na *Web*. Desde então, um número crescente de provedores de dados passaram a adotar estes princípios, conduzindo à criação de um espaço global de dados contendo bilhões de asserções sobre localizações geográficas, pessoas, companhias, livros, publicações científicas, filmes, música, programas de rádio e televisão, genes, proteínas, medicamentos e experimentos médicos, comunidades online, dados estatísticos, resultados de censos e publicações (BIZER, 2009).

3 Publicação de Linked Data

Publicar um conjunto de dados ligados na Web envolve, basicamente, as seguintes etapas (BIZER; HEATH; BERNERS-LEE, 2009):

1) Atribuir URIs às entidades descritas pelo conjunto de dados.
2) Definir ligações RDF para outras fontes de dados na Web, possibilitando navegar pela rede de dados, seguindo os links RDF.
3) Fornecer metadados sobre os dados publicados, de modo que se possa avaliar a qualidade dos dados publicados e escolher entre diferentes meios de acesso.

A atribuição de URIs às entidades consiste na definição de um identificador para cada uma delas(SAUERMANN; CYGANIAK, 2008). Em um ambiente aberto como a *Web*,

ISBN 978-85-61115-09-8

entretanto, é comum que diferentes provedores de informações publiquem dados sobre uma mesma entidade ou recurso (ex. uma localização geográfica ou uma celebridade), definindo diferentes URIs para identificar a mesma entidade. Como estas URIs referem-se à mesma entidade, são chamadas de *aliases* URIs (BIZER; HEATH; BERNERS-LEE, 2009), e têm uma importante função social na *Web* de Dados, visto que permitem diferentes visões e opiniões expressas na *Web*, sobre amesma entidade.

Para a publicação de *Linked Data* na Web, existem ferramentas que podem tanto servir conteúdos RDF, armazenados na forma de *Linked Data*, como prover visões de *Linked Data* a partir de bases de dados legadas não-RDF, facilitando o trabalho dos publicadores de conteúdo que não precisam lidar com detalhes técnicos, tais como negociação de conteúdo, provendo descrições RDF para as URIs (*deference*) e consultas na linguagem SPARQL, acrônimo para *SPARQL Protocol and RDF Query Language* (BIZER; CYGANIAK; HEATH, 2007; SAUERMANN; CYGANIAK, 2008;BIZER; HEATH; BERNERS-LEE, 2009).

Em *Linked Data*, a representação dos dados em RDF corresponde a um grafo, onde cada sequência *nó-arco-nó* é chamada de tripla, que representa uma relação *sujeito-predicado-objeto*, e são armazenadas em um tipo de banco de dados específico chamado *Triplestore* (ZHOU; WANG, 2009).

Da mesma forma como acontece com bancos de dados relacionais, a *Web* de Dados necessita de uma linguagem de consulta específica para RDF. A linguagem de consulta SPARQL possibilita que os consumidores de dados podem extrair informações possivelmente complexas, tais como referências a recursos e seus relacionamentos, que são retornadas, por exemplo, em um formato tabular, que pode incorporada em outras páginas *Web*.

A maior parte das consultas SPARQL contém um conjunto de padrões de triplas, onde sujeito, predicado e objeto podem ser variáveis. O padrão é comparado com o grafo RDF, e o sub-grafo correspondente aos critérios estabelecidos no padrão de triplas é o resultado da busca (PRUD'HOMMEAUX, 2008).

Há diversas alternativas de *Triplestores*RDF, como o *Virtuoso Universal Server*, Jena, Sesame, 4Store, YARS e OWLIM (VILLAZÓN-TERRAZAS *et al.*, 2011). Algumas destas já contam com recursos como *endpoint* SPARQL, que permite realizar consultas, e *Linked Data frontend*, que possibilita a navegação nos dados.

4 Contexto da Aplicação

O SINAES possui uma série de instrumentos complementares: auto-avaliação, avaliação externa, Exame Nacional de Desempenho de Estudantes(ENADE), Avaliação dos cursos de graduação e instrumentos de informação (censo e cadastro), que possibilita traçar um panorama da qualidade dos cursos e instituições de educação superior no País (INEP, 2013). Os resultados destas avaliações são públicos[21] e para serem utilizadas pelas Instituições de Ensino Superior (IES), para orientação da sua eficácia institucional e efetividade acadêmica e social; pelos órgãos governamentais para orientar políticas públicas; e pelos estudantes, pais de alunos, instituições acadêmico e público em geral, para orientar suas decisões quanto à realidade dos cursos e das instituições. Tais resultados são publicados e

[21]Exceção às informações de interesse privado das Instituições de Ensino Superior (IES), expressamente referidas na PORTARIA NORMATIVA Nº 40, DE 12 DE DEZEMBRO DE 2007 (dados relativos aos itens III, IV e X do art. 16, do Decreto nº 5773/ 2006). Exceção também para a divulgação dos resultados do Exame Nacional de Desempenho dos Estudantes – ENADE, onde é vedada a identificação nominal do resultado individual obtido pelo aluno examinado, que será a ele exclusivamente fornecido em documento específico, emitido pelo INEP.

ISBN 978-85-61115-09-8

disponibilizados a partir do portal do Ministério da Educação (MEC) e do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Também as IES, quando alcançam bons resultados, os divulgam e evidenciam em seus portais e campanhas de marketing.

A Web tem sido o principal meio, utilizado por governos, para a disseminação de informações digitais na internet. Contudo, Alves e Bax(2014) destacam alguns problemas: (1) dados da Administração Pública fragmentados, pouco integrados, em formatos distintos e, muitas vezes, proprietários e, portanto, de difícil reutilização (ALANI et al., 2007*apud* ALVES; BAX, 2014); (2) aplicações de exploração e visualização dos dados possibilitam baixa flexibilidade na sua manipulação.

Ou seja, esses dados permanecem ainda fechados e isolados, onde pouco ou nenhum compartilhamento e integração de conteúdos são praticados. A integração e reutilização desses dados ainda são difíceis, dada suas interfaces voltadas apenas para consulta ou extração ad-hoc, além dos altos custos e problemas envolvidos na análise de dados. São muitas as informações, que são publicadas - soltas na Web, sem significado bem definido, que são difíceis de serem interpretadas, transformando-se em um conteúdo de baixo valor agregado.

Se o cidadão quiser de fato participar e exercer o controle social sobre os processos de regulação, avaliação e supervisão da educação superior no sistema federal de educação, explorando a informação de forma diferente daquelas que foram previstas e implementadas pelos sistemas governamentais e portais das IES, enfrentará grandes dificuldades. Por exemplo, cruzar os dados dos portais MEC e INEP com os dados nos portais das IFES para aferir corpo docente, projeto pedagógico, matriz curricular, projetos, infraestrutura, etc., terá que fazer de forma não integrada e automatizada, cruzando dados muitas vezes não estruturados e alinhados.

Os trabalhos de Cabral *et al.*(2012) e de Oliveira e Turine (2012) apresentam abordagens que propõem a utilização de *Linked Data* na publicação dos dados do Exame Nacional do Ensino Médio (ENEM), como o objetivo de prover maior visibilidade e capacidade de análise sobre os dados relativos aos resultados da aplicação do exame.

Ramos e Marinho (2012) propuseram um modelo de informação para as Universidades Públicas Federais do Brasil a fim de ampliar a cultura de transparência e o nível de governança pública nos processos de regulação, avaliação e supervisão da educação superior no sistema federal de educação.Este modelo de informação especifica um conjunto de requisitos informacionais que devem compor, com outros requisitos específicos, a arquitetura de informação destas universidades. Serve de guia para a implementação dos portais web de acesso à informação das universidades federais e, principalmente, de um referencial, indicando quais as informações que deverão estar disponíveis a fim de viabilizar a participação do cidadão no processo de avaliação e regulação de tais instituições.

Uma vez que estas informações, especificadas por Ramos e Marinho (2012), sejam publicadas de forma ativa nos portais web destas universidades, com base nos princípios de dados abertos ligados - LOD, um conjunto de aplicações poderia ser desenvolvido, a fim de possibilitar a participação efetiva do cidadão nos processos de avaliação, supervisão e regulação da Educação Superior do Brasil. Atualmente, o acesso a estas informações não dispõe de uma aplicação e nem de um domínio LOD específico para isso, ou seja, uma base de dados abertas. As informações quando disponibilizadas ficam dispersas em vários ambientes e portais, o que dificulta muito o acesso por parte do cidadão, visto que muitos nem sabem que existem estas informações, onde estão disponíveis e tampouco como cruzar estes dados das IES com os dados do MEC/INEP.

Neste contexto, foi desenvolvido este protótipo de aplicação que tem por objetivo apresentar a integração de dados de avaliação, do MEC/INEP e os dados dos cursos

ISBN 978-85-61115-09-8

superiores das IFES do Brasil, levando em consideração a disponibilidade destes dados na *Web*. Os resultados são apresentados por meio do desenvolvimento de uma aplicação com as características de um *Mashup Website*, uma vez que consiste da união de diferentes fontes de dados e APIs (*Application Programming Interfaces*) em uma experiência de usuário integrada (ZANG, ROSSON e NASSER, 2008).

4.1 Obtenção, Tratamento e publicação dos dados

Para a demonstração da necessidade de padronização e integração, foram utilizados dados de quatro fontes diferentes. Além da origem, os dados estavam disponíveis em formatos diferentes, sendo necessário passarem por um processo de padronização e transformação para RDF. A Tabela 1 apresenta a origem e o formato dos dados utilizados.

Tabela 1- ***Fontes e formatos dos dados***

Conjunto de Dados	Fonte	Formato
Dados do ENADE 2012	INEP	CSV
Dados das IES	EMEC	*Screen Scrapping*
Localização dos Municípios	*Triplestore* LODKEM	RDF
Endereços Web de recursos dos cursos	Páginas dos cursos	Coleta manual

Fonte: elaborado pelos autores

Por se tratar de um estudo sobre a avaliação de cursos superiores, buscou-se o conjunto de dados do Exame Nacional de Desempenho de Estudantes (ENADE), que integra o SINAES, tendo como objetivo aferir o desempenho dos estudantes em relação aos conteúdos programáticos previstos nas diretrizes curriculares do respectivo curso de graduação, e as habilidades e competências em sua formação. O exame é realizado anualmente, e os dados mais recentes disponíveis na época deste estudo eram os da edição de 2012.

Os dados referentes ao ENADE 2012 foram obtidos a partir da página *Web* do INEP[22], em formato CSV, contendo os seguintes campos: Código e Descrição da Área (Curso); Código e Descrição da IES (Instituição); Categoria Administrativa; Organização Acadêmica; Código e Nome do Município do Curso; Código da UF; Número de Cursos na Unidade; Número de Estudantes Inscritos; Número de Estudantes Participantes; Nota Bruta do Curso – Formação Geral; Nota Padronizada do Curso – Formação Geral; Nota Bruta do Curso – Componente Específico; Nota padronizada do Curso – Componente Específico; Conceito Enade (contínuo); e, Conceito ENADE (faixa).

Visando complementar os dados referentes às instituições avaliadas, foram buscados na página do Sistema e-MEC[23], do Ministério da Educação. O e-MEC disponibiliza uma interface de consulta sobre Instituições e Cursos de Ensino Superior, conforme cadastro realizado pelas próprias instituições. A obtenção destes dados foi com da técnica de *Web scraping*, que permite a extração automatizada de pontos específicos de dados por meio da análise dos documentos codificados em linguagens de marcação, em vez de copiá-los manualmente (PENMAN, BALDWIN e MARTINEZ, 2009; VARGIU E URRU, 2013).

O *Web scraping* resultou na obtenção de dados de 2350 IFES brasileiras, públicas e privadas. Os dados extraídos são referentes à identificação, endereço das unidades, bem como contatos das pessoas responsáveis pelas instituições.

Com o objetivo de proporcionar uma visualização mais rica, buscou-se a inclusão do georeferenciamento das instituições, segundo as coordenadas de localização dos municípios

[22]http://portal.inep.gov.br/enade/resultados
[23]http://emec.mec.gov.br/

ISBN 978-85-61115-09-8

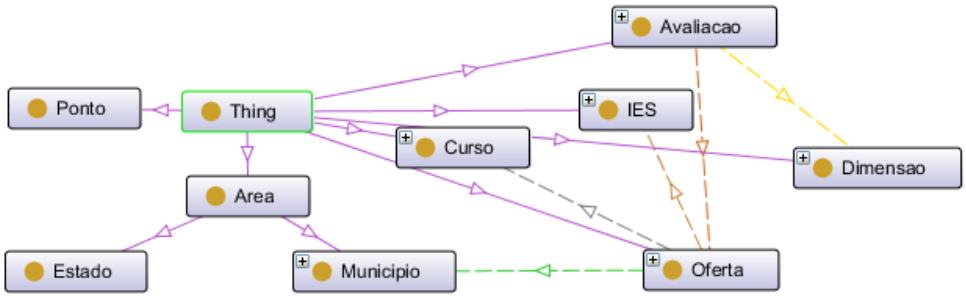
em que estas se encontram. Para isto, foram utilizados dados disponíveis em RDF no servidor do tipo *TripleStore* do LED-UFSC, que conta com vários outros conjuntos de dados associados aos municípios brasileiros (CABRAL *et al.*, 2012; DA SILVA *et al.*, 2013).

Finalmente, o quarto conjunto de dados utilizado diz respeito à disponibilidade de informações dos cursos avaliados em páginas *Web*. Para a obtenção destes dados, buscou-se identificar quais os cursos que disponibilizavam na *Web*: Página do Curso; Corpo Docente; Matriz Curricular; e, Projeto Político Pedagógico. Em um primeiro momento houve a tentativa de extração dos dados utilizando-se a técnica de *Web scraping*, mas, devido à falta de padronização dos *Websites* das instituições, passou-se para a coleta manual dos mesmos.

Para atender ao escopo desta aplicação, foi criada uma ontologia de domínio com o objetivo de representar os conceitos e a interligação dos dados. Em sua construção, foi utilizada, por meio de importação, a ontologia "geopoliticabr", pré-existente e disponível no âmbito do projeto Lodkem[24], e que representa as classes e propriedades referentes aos municípios brasileiros, e cujas instancias estão disponíveis na forma de *Linked Data* no *Triplestore* do projeto.

A Figura 1 apresenta as classes da ontologia e seus relacionamentos, incluindo as classes importadas. Com base na ontologia, o tratamento dos dados se consistiu na transformação dos diferentes formatos para o RDF. Para isto, foram utilizadas planilhas eletrônicas e *scripts* em linguagem de programação PHP, visando realizar a integração dos dados e sua ligação com os vocabulários utilizados.

Figura 1 - Ontologia de domínio de Avaliação



Fonte: elaborado pelos autores

Após a transformação dos dados para RDF, os mesmos foram publicados segundo os princípios de *Linked Data*, sendo armazenados no *Triplestore* do projeto LODKEM. Desta forma, os dados podem ser consultados em um *endpoint SPARQL*, e as URIs dos recursos podem ser acessadas por *dereference* em navegadores *Web.*

3.2 Demonstração dos resultados

A fim de demonstrar as potencialidades da interligação dos dados, esta seção descreve o desenvolvimento de uma aplicação do tipo *mashup website*que apresenta os dados previamente transformados e publicados como *Linked Data*. O objetivo da aplicação é apresentar, na forma de um painel, a visualização dos dados em diferentes formatos visuais, de forma a potencializar a capacidade de análise sobre os mesmos.

A aplicação foi desenvolvida utilizando as linguagens de programação PHP, com a biblioteca PHP-SPARQL-Lib, e *Javascript*, com as bibliotecas *Google Maps API*, *JQuery* e *Chart.js*.

[24] http://lodkem.ufsc.br/onto/geopoliticabr

ISBN 978-85-61115-09-8

Uma vez que os dados estão disponíveis em um servidor do tipo *triplestore*, as consultas necessárias são realizadas utilizando-se a linguagem SPARQL (PROD'HOMMEAUX e SEABORNE, 2008). O Quadro1 apresenta um trecho de consulta SPARQL onde é buscada uma lista com as informações dos cursos.

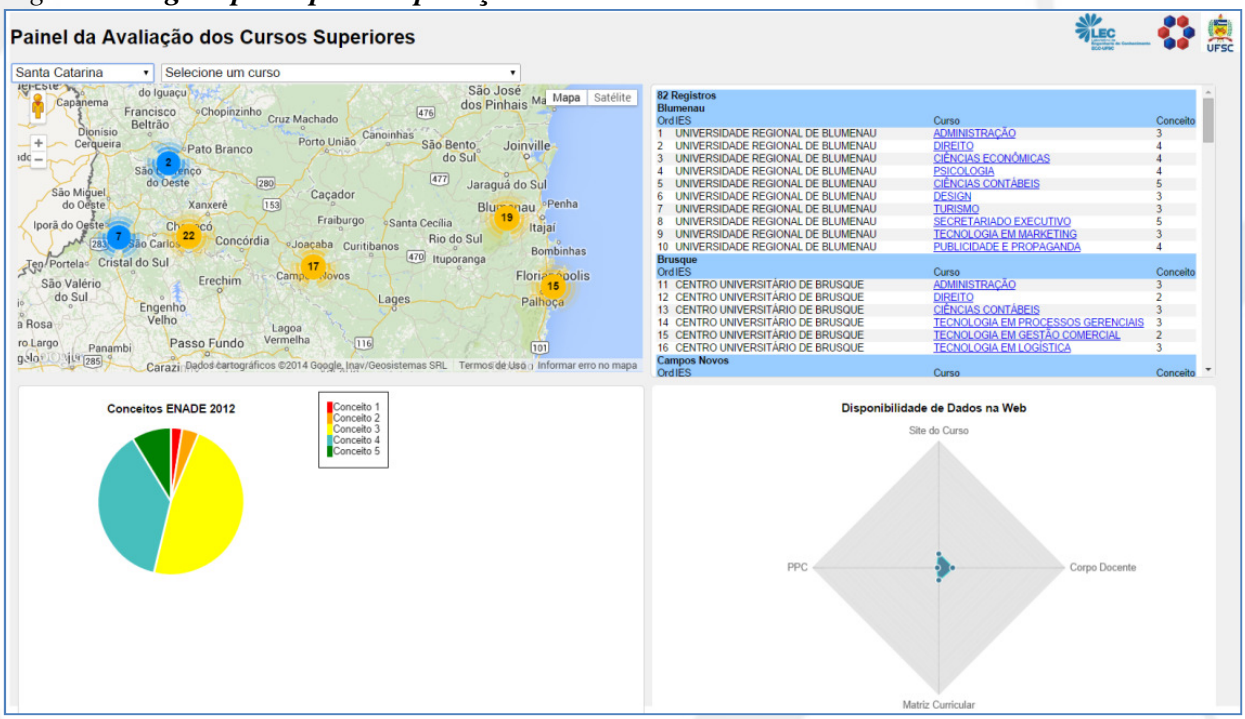
Quadro 1 – Trecho de consulta SPARQL

```
select ?uriCurso, ?codCurso, ?nomeCurso where{
  ?uriCurso a <http://lodkem.ufsc.br/onto/avaliacaoSuperior#Curso> .
?uriCurso <http://lodkem.ufsc.br/onto/avaliacaoSuperior#temCodCurso> ?codCurso .
  ?uriCurso rdfs:label ?nomeCurso .
}
order by ?nomeCurso
```

Fonte: elaborado pelos autores

A aplicação desenvolvida conta com uma página principal desenvolvida utilizando bibliotecas *javascript*, que fazem requisições a scripts desenvolvidos em PHP, responsáveis por montar e realizar as consultas ao *endpoint SPARQL*. Os resultados das consultas são, então, serializados em JSON (*Javascript Object Notation*) e devolvidos à página solicitante. A Figura 2 apresenta a página principal da aplicação, onde é possível observar, na parte superior, as caixas de seleção para filtragem dos dados por estados e/ou cursos.

*Figura 2–**Página principal da aplicação***



Fonte: elaborado pelos autores

O primeiro tipo de visualização disponível é um mapa, onde as instituições são apresentadas segundo a localização dos municípios em que se encontram.

O segundo tipo de visualização é uma lista das instituições e cursos, ordenados por município. Nesta lista, os nomes dos cursos são *hyperlinks* para uma nova janela onde são apresentados os dados específicos daquele curso selecionado.

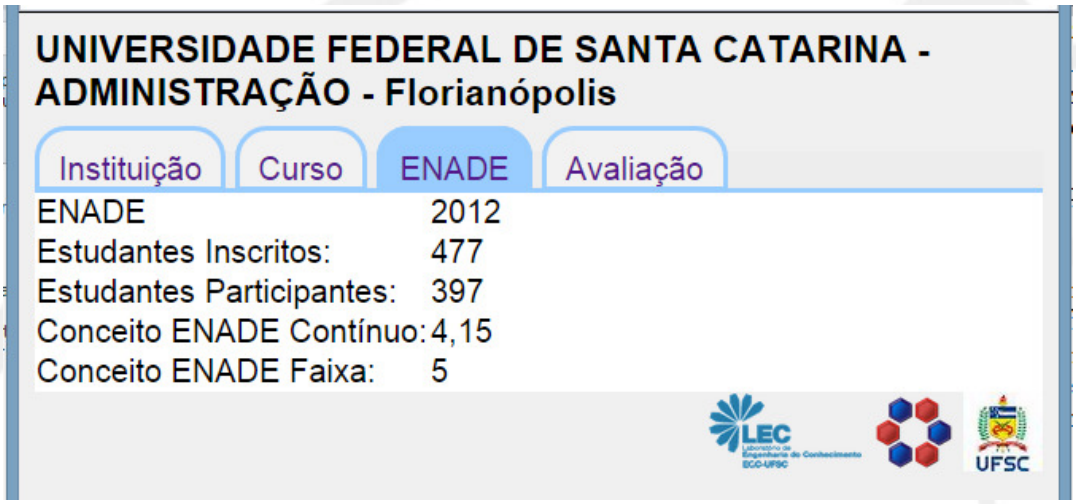
A terceira visualização disponível na página é um gráfico de pizza onde são exibidas as porcentagens dos cursos selecionados segundo os conceitos obtidos no ENADE 2012.

ISBN 978-85-61115-09-8

Finalmente, o quarto tipo de visualização de dados é um gráfico do tipo radar, onde são exibidas as porcentagens da disponibilidade de dados na *Web* dos cursos selecionados. São levados em consideração, neste gráfico, a disponibilidade de *website* do curso, corpo docente, matriz curricular e Projeto Pedagógico do curso.

Além da visualização dos dados agrupados, disponível na página principal, as informações específicas de cada curso podem ser visualizadas clicando no *hyperlink* correspondente.

Figura 3 - ***Janela de detalhes do curso***



Fonte: elaborado pelos autores

A Figura 3apresenta a janela que traz as informações específicas de cada curso, divididas em quatro abas: Instituição, que apresenta as informações referentes à IES; Curso, que apresenta as informações específicas do curso selecionado; ENADE, que apresenta as informações relativas ao ENADE; e, Avaliação, que apresenta as informações sobre a disponibilidade na *Web* de seu site, corpo docente, matriz curricular e Projeto Pedagógico do Curso.

Considerações Finais

Este trabalho englobou temas extremamente atuais, importantes, relevantes e complexos para sociedade brasileira: *a Lei de Acesso à Informação, Participação e Controle Social, Dados Abertos e a Regulação, Avaliação e Supervisão da Educação Superior no Brasil*. No que tange à Educação Superior no Brasil, a participação da sociedade como um todo é fundamental para garantia da qualidade e da transparência de modo a atender aos anseios da sociedade.

Para que realmente a sociedade, como um todo, possa contribuir para a melhoria da qualidade do ensino superior faz-se necessário além do amplo acesso as informações dos processos de regulação, avaliação e supervisão das IFES brasileiras, mas, principalmente, que sejam desenvolvidas ferramentas, que permitam a efetiva participação do cidadão, de forma individual ou por meio da sociedade civil organizada.

Atualmente, o acesso a estas informações não se dá por meio de uma aplicação integradora e/ou a um domínio LOD específico. As informações quando disponibilizadas ficam dispersas em vários ambientes e portais, o que dificulta muito o acesso por parte do

ISBN 978-85-61115-09-8

cidadão comum, visto que muitos nem sabem que existem estas informações e tampouco onde estão disponíveis.

A partir destes requisitos e dos princípios de Web Semântica, *Linked Data* e do Modelo de Informações, descrito por Ramos e Marinho (2012), foi desenvolvida uma aplicação *Web mashup* para a visualização dos dados com auxílio de consultas SPARQL.O aplicativo cruza os dados dos portais MEC e INEP com os dados nos portais das IFES para aferir corpo docente, projeto pedagógico, matriz curricular, resultados ENADE, informações gerais, de forma integrada e automatizada.

Com efeito, esta aplicação demonstrou o potencial de ferramentas desta natureza, visto quesão ampliadas a visão, a compreensão e a cultura de governança pública, onde o papel do cidadão nesta política pública assume outra dimensão a partir da viabilidade de sua participação e controle social sobre a educação superior no Brasil.

Por fim, conclui-se que para a efetiva participação dos cidadãos no sistema de nacional de avaliação do ensino superior é necessário que não só o Estado, mas principalmente as IFES disponibilizem os dados em formato aberto para o desenvolvimento de novas aplicações e o consequente aprimoramentos dos mecanismos de governança.

Como trabalhos futuros, sugere-se a inclusão dos demais dados referentes aos cursos e suas avaliações disponíveis no Sistema e-MEC, bem como dos dados das outras edições do ENADE, possibilitando a análise da evolução histórica das avaliações, agregando valor às informações. Outro ponto importante diz respeito à periodicidade de atualização dos dados. Para isso, sugere-se a construção de um extrator automatizado para os dados do Sistema e-MEC que realize a exportação para o formato RDF, e atualização dos mesmos no *Triplestore*.

Ainda, considerando-se a pouca disponibilidade de dados nos *websites* dos cursos, bem como a falta de um padrão dos mesmos, que dificulta a extração automatizada por meio da técnica de *Web scraping*, sugere-se a criação de um serviço para preenchimento dos dados por parte dos usuários, utilizando ferramentas de *Web* 2.0, garantindo, assim, a participação da sociedade no monitoramento e atualização dos dados.

Referências

ALVES, Marcus Vinícius Chevitarese;BAX, Marcello Peixoto. Da necessidade e viabilidade da adoção do padrão LOD pelo Congresso Nacional: um estudo no contexto do orçamento público. **Informação e Sociedade**: **Estudos**, João Pessoa, v.24, n.1, p. 73-94, jan./abr. 2014.

BERNERS-LEE, Tim. **Information Management: A Proposal**. 1989. Disponível em: http://www.w3.org/History/1989/proposal.html. Acesso em: 25 fev. 2014.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The Semantic Web. **Scientific American**. 2001. Disponível em: <http://www.scientificamerican.com/article/the-semantic-web>. Acesso em 10 ago. 2014.

BERNERS-LEE, Tim. **Linked Data -Design Issues**. 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em 12 ago. 2014.

BIZER, Chris; CYGANIAK, Richard; HEATH, Tom. How to Publish Linked Data on the Web. **7th International Semantic Web Conference (ISWC2008)**. Karlsruhe, Alemanha, 2008.

BIZER, Chris. The Emerging Web of Linked Data. **IEEE Intelligent Systems**. Vol 24, No. 5. 2009.

BIZER, Chris; HEATH, Tom; BERNERS-LEE, Tim. Linked Data - The Story So Far. **International Journal on Semantic Web and Information Systems**, Vol. 5(3), Pages 1-22. 2009. DOI: 10.4018/jswis.2009081901

ISBN 978-85-61115-09-8

BRADLEY, Fiona. Discovering Linked Data. **Library Journal**. Vol 134, No. 7. 2009.

BRASIL. **Lei nº. 10.861, de 14 de abril de 2004**, 2004 Disponível em: <http://www.mec.gov.br>. Acesso em: 18 mai. 2012.

BRASIL. **Lei nº. 12.527, de 18 de novembro de 2011 - Lei de Acesso à Informação**, 2011 Disponível em:<http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 10 mai. 2012.

BRASIL. **Decreto nº 8.243, de 23 de maio de 2014 - Política Nacional de Participação Social**, 2014. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2014/Decreto/D8243.htm>. Acesso em: 17 mai. 2014.

CABRAL, Samuel Pierri; BEDUSCHI, Nitay Batista; ZANCANARO, Airto; TODESCO, José Leomar; GAUTHIER, Fernando A. O. Aplicando Linked Data na publicação de dados do ENEM. **Proceedings of Joint V Seminar on Ontology Research in Brazil and VII International Workshop on Metamodels, Ontologies and Semantic Technologies**. Recife, Brasil, 2012.

DA SILVA, Gesiel; FRANZEN, Greici Barreta; TODESCO, José Leomar; GAUTHIER, Fernando A. O.; SPERONI, Rafael de Moura. Processo de publicação de linked data para a distribuição de informação e conhecimento. **III Congresso Internacional do Conhecimento e Inovação**. Porto Alegre, Brasil. 2013

HEATH, Tom. Welcome to the Data network. **IEEE International Computing**. Vol 15, No. 6. 2011.

HEATH, Tom; BIZER, Christian. Linked Data: Evolving the Web into a Global Data Space (1st edition). **Synthesis Lectures on the Semantic Web: Theory and Technology**, 1:1, 1-136. Morgan & Claypool. 2011.

HYLAND, Bernadette; WOOD, David. The Joy of Data – A Cookbook for Publishing Linking Gonvernment Data on the Web. In: **Linking Government Data**. Springer. USA, 2010.

INEP. **Sinaes**. Disponível em:<http://portal.inep.gov.br/superior-sinaes>. Acesso em: 17 mar. 2013.

JACOBS, Ian; WALSH, Norman. **Architecture of the World Wide Web**, Volume One. W3C Recommendation. 2004.

MEC. **Portaria Nº 1.264**, De 17 de outubro de 2008. Disponível em: <http://meclegis.mec.gov.br>. Acesso em: 07 mai. 2012.

MEC. **Portaria Nº 40**, De 12 de dezembro de 2007 - Republicada em 29 de dezembro de 2010. Disponível em: <http://meclegis.mec.gov.br>. Acesso em: 05 mai. 2012.

MEC. **Projeto de lei que aprova o Plano Nacional de Educação 2011-2020**, 2012. Disponível em:<http://conae.mec.gov.br/index.php?option=com_content&view=article&id=363:pne&>. Acesso em: 05 jul. 2012.

OLIVEIRA, Davison A. Z.; TURINE, Marcelo A. S. **Uma estratégia para Publicação de um *Linked Open Data* Baseado em *Data Warehouse***. Dissertação de Mestrado em Ciência da Computação. Universidade Federal do Mato Grosso do Sul. 2012.

OPEN KNOWLEDGE FOUNDATION. **Open Data Handbook Documentation**. Open Data Handbook. 2012. Disponível em : <http://opendatahandbook.org>. Acesso em 10 jun. 2014.

ISBN 978-85-61115-09-8

PENMAN, Richard Baron; BALDWIN, Timothy; MARTINEZ, David. **Web Scraping Made Simple with SiteScraper**. 2009 Disponível em: <https://code.google.com/p/sitescraper>

PROD'HOMMEAUX, Eric; SEABORNE, Andy. **SPARQL Query Language for RDF**. W3C Recommendation. 2008. Disponível em: <http://www.w3.org/TR/rdf-sparql-query/> Acesso em 20 ago. 2014

RAMOS, Alexandre Moraes; MARINHO, Sidnei Vieira. **Modelo de informação para divulgação das informações dos processos de avaliação e regulação das universidades públicas federais no contexto da lei de acesso à informação**. 2012. Trabalho de Conclusão de Curso de Bacharel em Administração – Universidade Federal de Santa Catarina, Florianópolis, 2012.

SAUERMANN, Leo; CYGANIAK, Richard. **Cool URIs for the Semantic Web**. World Wide Web Consortium. Notes. 2008. Disponível em: <http://www.w3.org/TR/cooluris/> Acesso em 04 jul. 2014.

SCHLICKMANN, R.; MELO, P. A.; ALPERSTEDT, G. D. Enfoques da teoria institucional nos modelos de avaliação institucional brasileiros. **Avaliação**, Campinas; Sorocaba, SP, v. 13, n. 1, p. 153-168, mar. 2008.

SOUZA, I. M. **Gestão das universidades federais brasileiras:** uma abordagem fundamentada na gestão do conhecimento. Tese (Doutorado) - Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2009.

VARGIU, Eloisa; URRU, Mirko. Exploiting web scraping in a collaborative filtering-based approach to web advertising. **Artificial Intelligence Research**. Vol. 2, No. 1. 2013

VILLAZÓN-TERRAZAS, Boris; VILCHES-BLÁSQUEZ, Luiz M.; CORCHO, Oscar; GÓMEZ-PÉREZ; Asunción. Methodologial Guidelines for Publishing Gonvernment Linked Data. **Linking Government Data**. Springer, USA. 2010.

ZANG, N., ROSSON, M.B. & NASSER, V. Mashups: Who, What, and Why? Extended Abstracts of **Human Factors in Computing Systems: CHI 2008**. 2008.

ZHOU, Jing; WANG, Kemin. Towards Scalable Decentralized TripleStores. **International Symposium on Computer Network and Multimedia Technology**. Wuhan, China. 2009.

ISBN 978-85-61115-09-8

VISUALIZAÇÃO DE DADOS ABERTOS VINCULADOS EM SISTEMAS DE INFORMAÇÕES GEOGRÁFICAS: UMA REVISÃO SISTEMÁTICA DA LITERATURA

Patricia Carolina Neves Azevedo
paty.neves@gmail.com
Júlia Epischina Engrácia de Oliveira
juliae@fumec.br
Fernando Silva Parreiras
fernando.parreiras@fumec.br

Resumo

A apresentação visual dos dados pode facilitar a compreensão de forma intuitiva. Com o crescente volume de dados publicados na web, os Sistemas de Informação Geográfica surgem como ferramenta para visualização de dados abertos vinculados. Entretanto, ainda faltam trabalhos que apresentem uma visão sistemática sobre o estado-da-arte das pesquisas neste campo de conhecimento. Este trabalho objetiva caracterizar as publicações relacionadas à visualização de *linked data* em SIGs por meio de uma revisão sistemática da literatura. Foram selecionadas 55 publicações na área, após aplicação dos critérios de exclusão, para análise das tendências na literatura científica. Conclui-se que as pesquisas, em sua maioria, se concentram a partir de 2010, utilizam dados governamentais, são avaliadas por meio de exemplos e descrevem métodos ou meios de desenvolvimento.

Palavras-chave: Linked Open Data. Revisão Sistematica da Literatura. Sistemas de Informação Geográfica.

Abstract

As the amount of structured data published on the Web using open standards increase, Geographical Information Systems emerge as an alternative to visualize linked open data. However, studies providing a systematic analysis of this field lack so far. In this paper, we describe the results of a systematic mapping study on the applications of linked data in geographical information systems. After applying the exclusion criteria, we classify 55 papers and conclude that most of the research on this topic was published after 2010, using government data, describes methods techniques and is validated with examples.

Key Words: Linked Open data. Systematic Literature Review. Geographic Information Systems.

Introdução

A apresentação visual dos dados de forma clara permite que os usuários possam obter uma melhor compreensão dos mesmos. A literatura aponta os Sistemas de Informação Geográfica (SIG) como importante ferramenta para visualização de *linked data*, por

ISBN 978-85-61115-09-8

permitirem a integração de dados vindos de várias fontes heterogêneas, com o propósito de potenciar a descoberta e a divulgação de novos conhecimentos. Este trabalho objetiva caracterizar as publicações relacionadas à visualização de *linked data* em SIGs por meio de uma revisão sistemática da literatura.

A revisão sistemática da literatura, ou SLR, constitui um instrumento fundamental para a realização de pesquisas baseada em evidências. A intenção de uma SLR é abordar uma questão de pesquisa bastante específica, a fim de determinar onde existem agrupamento de estudos que poderiam apoiar uma análise mais completa. A SLR envolve uma pesquisa da literatura capaz de determinar os tipos de estudos sobre visualização de *linked data* e SIG.

No entanto, como normalmente ocorre na engenharia de software, o desenvolvimento aparenta ser guiado mais por opinião de especialistas que baseado em evidências empíricas ou modelos cognitivos, levando a dúvidas sobre o que realmente se sabe, os benefícios e as limitações do objeto de pesquisa.

Esta revisão sistemática da literatura será capaz de sintetizar e apresentar as buscas empíricas sobre visualização de *linked data* em SIGs, assim como uma visão geral sobre o estado da arte, que acreditamos ser importante para a comunidade científica para construir um entendimento comum dos desafios que devem ser enfrentados sobre os tópicos abordados.

O artigo está organizado da seguinte forma: na seção 2 tem-se uma visão geral sobre visualização de dados geoespaciais e *linked data*. A seção 3 conceitua a revisão sistemática da literatura, descreve os métodos utilizados para esta revisão, apresenta os resultados divididos em tipos de pesquisa e análise temporal e menciona as limitações da pesquisa. A seção 4 conclui a pesquisa e sugere recomendações para outras pesquisas sobre visualização de *linked data* em SIGs.

1 Visualização de Dados

Steele e Iliinsky (2011) dissertam sobre a visualização de dados como um eficiente e eficaz meio de comunicação para um grande volume de informações, e conceituam que os termos de visualização de dados e de visualização de informações são úteis para se referir a qualquer representação visual dos dados, que são:

a) algoritmicamente desenhados (podem ter toques personalizados, mas é amplamente renderizado com a ajuda de métodos computadorizados);

b) fácil de se regenerar com dados diferentes (o mesmo formulário pode ser reaproveitado para representar conjuntos de dados diferentes, com dimensões ou características semelhantes);

c) muitas vezes esteticamente árido (dados não decorados);

d) relativamente rico em dados (grandes volumes de dados são bem vindos e viáveis).

Visualizações de dados são inicialmente projetados por um humano, mas são desenhadas graficamente por algoritmos ou software de diagramação. A vantagem dessa abordagem é o fato de ser relativamente simples para atualizar ou gerar novamente a visualização incluindo novos dados (Steele e Iliinsky, 2011).

A visualização dinâmica de dados é uma das formas culturais genuinamente nova que se tornou possível graças à computação (Manovich, 2009). Manovich analisa a visualização de um conjunto de dados com computadores como uma possibilidade mais ampla, capaz de

ISBN 978-85-61115-09-8

criar visualizações dinâmicas, alimentar dados em tempo real, basear as representações gráficas de dados em sua análise matemática usando vários métodos, da estatística clássica à prospecção de dados, mapear um tipo de representação em outro (imagens em sons, sons em espaços tridimensionais, etc.).

Os autores Steele e Iliinsky (2011) são explícitos sobre o motivo da visualização ser um meio útil para examinar, compreender e transmitir a informação:

a) visualização aproveita as capacidades incríveis do sistema visual para mover uma enorme quantidade de informações para o cérebro muito rapidamente;

b) visualização permite identificar padrões, relacionamentos e seus significados;

c) visualização ajuda a identificar subproblemas;

d) visualização é realmente bom para identificação de tendências ou produtos fora de série, descobrindo pontos específicos ou interessantes em um campo maior, etc.

1.1 Visualização de Dados Geoespaciais

A informação geográfica se distingue de outras informações por referir-se a objetos ou fenômenos com uma localização específica no espaço e, portanto, tem um endereço espacial (Kraak e Ormeling, 2003). Os mesmos autores explicam que devido a essa característica, os locais dos objetos ou fenômenos podem ser visualizados, e essas visualizações, chamadas de mapas, mostram como os objetos do mundo real, por exemplo: como casas, estradas, campos ou montanhas podem ser abstraídos como um modelo de paisagem digital, de acordo com alguns critérios pré-determinados, e armazenados em SIGs, entre eles pontos, linhas, áreas ou volumes. Quando armazenados em um banco de dados, Kraak e Ormeling (2003) dividiram esses dados geoespaciais em dados de localização, dados de atributos e dados temporais:

a) componentes localização, atributo e tempo e suas perguntas relacionadas: onde, o quê e quando;

b) a visualização do objeto;

c) características detalhadas dos componentes dos dados;

Kraak e Ormeling (2003) justificam a unicidade de um SIG pela capacidade de combinar dados geoespaciais e não geoespaciais de diferentes fontes de dados em uma operação de análise geoespacial, a fim de responder a vários tipos de perguntas.

O desenvolvimento de SIGs foi estimulado por áreas individuais, tais como a defesa civil, cadastros, serviços públicos e planejamento regional. Já que todas as áreas têm origens e necessidades diferentes, a funcionalidade do software SIG se torna diferente a cada tipo de necessidade (Kraak e Ormeling, 2003).

Os SIGs são eficientes na combinação de conjuntos de dados, não obstante o fato desses dados serem de épocas e resoluções diferentes, ou até não passíveis de combinação, o software combina esses dados e apresenta os resultados (Kraak e Ormeling, 2003).

1.2 Visualização de *Linked Data*

A visualização e a interação de *linked data* é uma questão que tem sido reconhecida desde o início da web semântica, Geroimenko e Chen (2003). Ao aplicar técnicas de visualização de informação, a web semântica auxilia os usuários na exploração e interação

ISBN 978-85-61115-09-8

dos dados. A transformação e apresentação visual desses dados são os principais objetivos da visualização de informação, de tal modo que os usuários possam obter uma melhor compreensão dos dados (Card et al., 1999). Visualizações são úteis para a obtenção de uma visão geral dos *datasets*, seus tipos principais e as relações entre eles.

A visualização de dados pode ser definida como algo que dá ao usuário uma maneira de analisar os dados, de modo a obter conhecimento e entendimento. Já a visualização de dados vinculados é uma exibição de dados que se comunica com outra visão. Se uma modificação é feita para uma das visões, o outro ponto de vista vai mudar sua aparência em reação àquela modificação (Chen et al., 2007).

Visualização de dados ligados se enquadra na categoria de navegação baseada em ontologia em busca de informações, onde anotação semântica de dados é utilizada para apoiar a exploração desses dados (Paulheim e Probst, 2010).

Para uma utilização eficaz, é essencial fornecer mecanismos simples para consultar os *datasets.* Ahlberg et al. (1992) conceituam consultas dinâmicas como sendo a interface gráfica com manipulação direta, como por exemplo, listas ou *slide-bars* que, quando alterados, consultam automaticamente o banco de dados e os dados do filtro são exibidos. Shneiderman (1996) explica que, primeiramente, todos os dados são exibidos, então o usuário utiliza os filtros para selecionar o subconjunto de interesse, e será visualizado os detalhes destes novos dados.

2 Revisão Sistemática da Literatura

O paradigma baseado em evidências é amplamente utilizado na medicina clínica e na educação, como uma ferramenta para apoiar a prática e formulação de políticas. O conceito básico que sustenta esta técnica é a realização de um estudo secundário que sistematicamente localiza, avalia e agrega os resultados de um conjunto de estudos empíricos, a fim de reunir as melhores evidências disponíveis para responder a uma pergunta de pesquisa de forma imparcial. A ideia de adaptá-lo para uso em engenharia de software foi proposta pela primeira vez em 2004 por Kitchenham et al. e, desde então, tornou-se cada vez mais aceito como um complemento útil para o conjunto de ferramentas metodológicas utilizada na engenharia de software.

A revisão sistemática da literatura é um método importante para resumir e fornecer uma visão geral da maturidade da disciplina (Kitchenham et al., 2004), que busca um sentido em grandes volumes de informação e um meio de contribuição para as respostas às questões sobre o que funciona e o que não funciona - entre vários outros tipos de perguntas. É um método de mapeamento e identificação de áreas de incerteza e onde ainda são necessários estudos, por não ter nenhuma ou pouca pesquisa relevante sobre o assunto. A revisão sistemática também sinaliza áreas onde existem falsas certezas, áreas onde pensamos que sabemos mais do que realmente sabemos e que, na verdade, existem poucas evidências para apoiar essas crenças (Petticrew e Roberts, 2006).

A SLR é definida por um protocolo que estabelece as etapas dos procedimentos a serem realizados durante a revisão. Os procedimentos metodológicos, presentes no protocolo,

ISBN 978-85-61115-09-8

representam as "forças" da SLR, permitindo tanto avaliar o estado atual dos conhecimentos da área, como manter a atualização de pesquisas em base avançada (Cook et al., 1997). Uma SLR difere de uma revisão da literatura simples ou de um survey por ser um estudo replicável, científico e transparente, evitando assim os vieses.

O objetivo desse método é fornecer uma oportunidade alternativa de melhor visualização do contexto da pesquisa em questão, combinando e analisando resultados quantitativos de estudos empíricos, a fim de dar sentido à literatura em constante evolução (Glass, 1976).

Diante da literatura em crescimento, cujo conhecimento encontra-se inexplorado, a SLR merece maior prioridade que a adição de um novo experimento ou survey (Glass, 1976). O acúmulo de conhecimento depende cada vez mais da integração entre estudos anteriores e descobertas empíricas (King e He, 2005).

2.1 Planejamento

Este estudo foi realizado como uma revisão sistemática da literatura (tradução de Systematic Literature Review - SLR) com base nas diretrizes originais propostas por Kitchenham (Kitchenham et al., 2009) e com o propósito de responder às seguintes perguntas:

Q1. Quais são os tipos de pesquisa mais utilizados ao se tratar de visualização em SIG e *linked data*?
Q2. A partir de 2010, qual a frequência do uso de dados governamentais em pesquisas que relacionam visualização em SIG e *linked data*?
Q3. Quais foram os tipos de resultado obtidos com o uso de dados governamentais?

2.2 Realização

Na pesquisa, utilizou-se o Google Scholar por ser um motor de busca em bases de dados confiáveis, de documentos, artigos científicos, revisões, papers de conferências, repositórios de documentos digitais, institucionais e multidisciplinares, reconhecidos pela comunidade acadêmica internacional.

A definição dos termos para as buscas procedeu-se através da combinação das seguintes palavras-chave: *linked, data*, visualization, geovisualization, maps, geographic information system, gis, semantic, web, spatial, geographic, *datasets*. Utilizando os operadores booleanos OR e AND, foram feitas combinações de termos para formação da string de pesquisa, segundo a Tabela 1.

Quadro 1 -*Strings* da pesquisa

STRINGS DA PESQUISA
("geographic information system" OR gis) AND (visualization OR geovisualization OR "data visualization") AND ("web semantic" OR semantic) AND ("linked data")

ISBN 978-85-61115-09-8

A pesquisa realizada nas bases de dados permitiu a seleção de 55 publicações após a eliminação de 58 publicações, seguindo os seguintes critérios de exclusão:

- monografias, editoriais, prefácios, sumários, entrevistas, notícias, revisões, tutoriais, workshops, painéis e pôster;
- publicações que não estejam em inglês ou português;
- publicações pagas.

Na etapa seguinte da revisão sistemática da literatura, foi feita a leitura e análise dos textos completos das publicações selecionadas para classificá-las de acordo com o tipo de publicação, resultado da pesquisa e ano de publicação.

2.3 Resultados

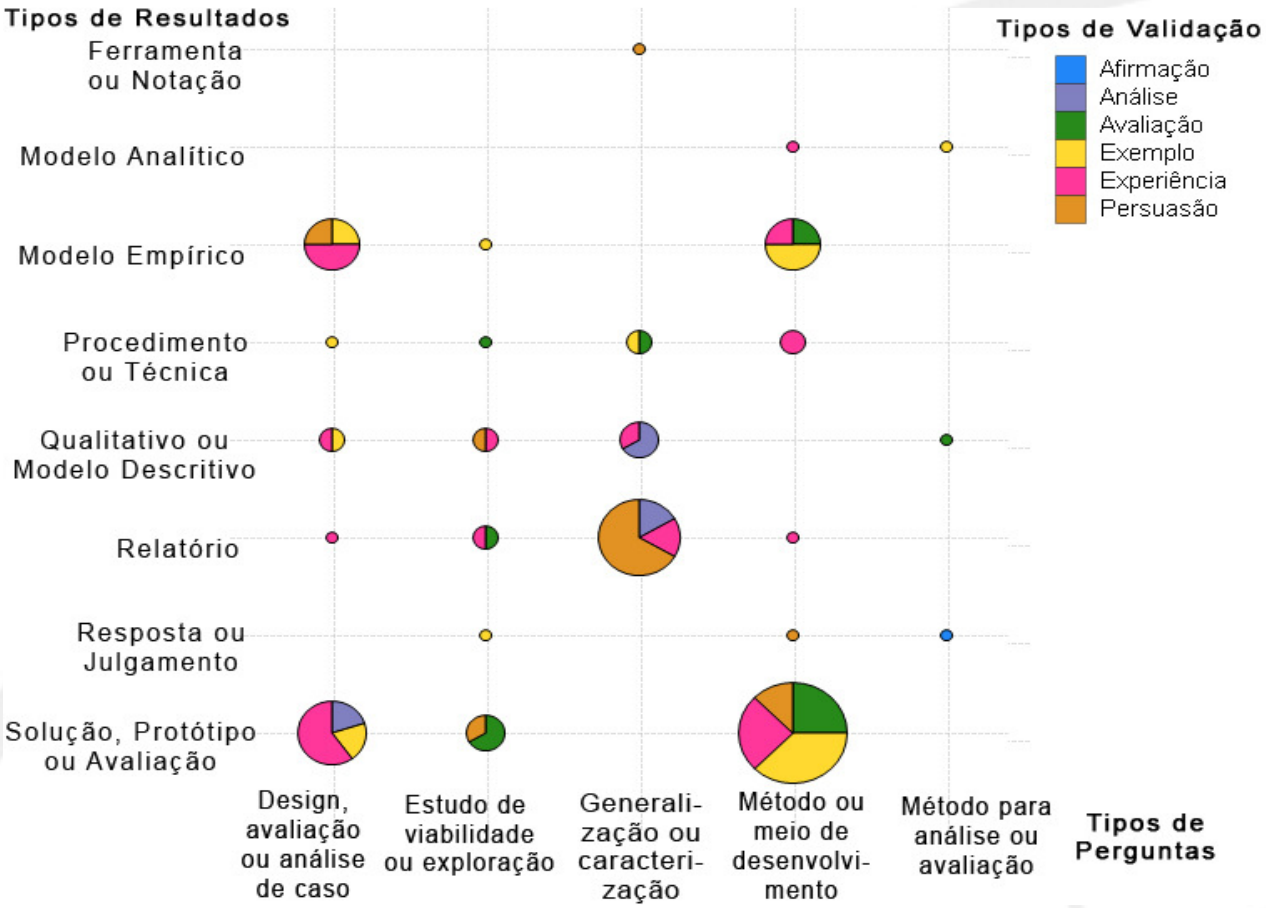
2.3.1 Q1: Tipos de Pesquisas

A Figura 1 apresenta um modelo que explica trabalhos de pesquisa em aplicações que envolvem *linked data*, classificando em três níveis: os tipos de questões de investigação que solicitam, os tipos de resultados que produzem e o caráter da validação que fornecem. Este modelo pertence à engenharia de software e vem evoluindo ao longo de vários anos, desde a versão apresentada inicialmente por Mary Shaw, na ICSE (International Conference on Software Engineering) em 2001.

As pesquisas em engenharia de software respondem a perguntas sobre métodos de desenvolvimento ou análise, sobre detalhes do projeto ou avaliação de um caso particular, sobre generalizações, classes de sistemas ou técnicas, ou sobre questões exploratórias visando existência ou a viabilidade de uma tarefa (Shaw, 2002).

ISBN 978-85-61115-09-8

Figura 1 - Tipos de Pesquisas em Aplicações *Linked Data*.



As contribuições tangíveis nas pesquisas em engenharia de software podem ser procedimentos ou técnicas para o desenvolvimento ou análise, podem ser modelos que generalizam a partir de exemplos ou podem ser ferramentas específicas, soluções ou resultados sobre sistemas particulares (Shaw, 2002).

O último nível do modelo descreve os tipos de validação para suportar os resultados da pesquisa. É essencial selecionar a forma de validação apropriada para o tipo de resultado e o método utilizado para obter o resultado (Shaw, 2002).

A Figura 1 exibe a resposta da questão Q1, onde visualiza-se que a combinação mais utilizada nos tipos de pesquisa que tratam de visualização em SIG e *linked data* foram perguntas sobre o método ou meio de desenvolvimento; soluções, protótipos ou avaliações como resultado; e exemplos como forma de validação. Neste caso, há a tendência em saber como criar ou automatizar e qual o melhor jeito de fazê-lo, sendo testado por meio de um sistema que, em execução, incorpore ou seja portador do resultado, ou ainda, que a sua implementação ilustre um princípio que pode ser aplicado em outros lugares. O uso de exemplo é adequado a esta combinação, sendo uma evidência convincente da validação do resultado obtido, como um sistema desenvolvido.

Pode-se observar também que a maioria dos relatórios que respondem a perguntas sobre generalização ou caracterização utilizam a persuasão como forma de validação. Neste cenário, a validação puramente pela persuasão raramente é suficiente para um trabalho de

ISBN 978-85-61115-09-8

pesquisa. Porém, se a pergunta original for sobre viabilidade, um sistema em funcionamento, mesmo sem análise, pode ser suficiente (Shaw, 2002). Ao verificar este novo cenário na Figura 1, conclui-se que menos da metade das publicações que estudam a viabilidade e resultam em um sistema em funcionamento, utilizam a persuasão como forma de validação.

Quanto aos tipos de resultados, prevaleceram aqueles que abordavam uma solução de aplicação para um problema que demonstra o uso dos princípios da engenharia de software. Os tipos de perguntas que mais têm sido exploradas são sobre método ou meio de desenvolvimento. Os métodos para análise ou avaliação são os tipos de perguntas menos explorados. Em relação ao tipo de validação, as soluções apresentadas têm explorado mais a experiência, seguida de exemplo e persuasão. Somente um artigo abordou a afirmação como forma de validação e deve ser visto como um ponto positivo, já que neste caso nenhuma tentativa séria é utilizada para avaliar o resultado.

Percebe-se uma lacuna nas pesquisas que geram uma ferramenta ou notação. Somente um artigo foi caracterizado como uma nova ou melhor maneira de fazer alguma tarefa, medição técnica ou avaliação, incluindo técnicas operacionais para execução, representação, gestão e análise, mas excluindo os que recomendam diretrizes.

2.3.2 Q2: Uso de Dados Governamentais

A Figura 2 ilustra a combinação entre tipos de resultados, ano das pesquisas e se estas fizeram uso de dados governamentais. Ao observar esta figura, é possível notar que a partir de 2010 houve um crescimento no interesse em uso de dados governamentais nas pesquisas aplicadas à *linked data*. Esse ponto tende a aumentar com a influência da Lei nº 12.527, a Lei de Acesso à Informação, sancionada em 18/11/2011 e em vigor em 16/05/2012. De acordo com o seu regulamento, "é dever dos órgãos e entidades públicas promover, independentemente de requerimentos, a divulgação em local de fácil acesso, no âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas".

ISBN 978-85-61115-09-8

Figura 2 - Características das Pesquisas sobre Visualização em SIG e *Linked Data.*

A Figura 2 responde a questão Q2, ao apontar que grande parte das publicações sobre a temática enfocada na pesquisa foram em 2010, 2011 e 2012 com 16, 13 e 19 artigos respectivamente, sendo possível analisar o desenvolvimento do enfoque da pesquisa no decorrer do tempo, as características, resultados e utilização de conhecimentos acadêmicos e científicos produzidos por diversos pesquisadores.

Como é possível perceber, o estudo gera várias possibilidades de futuras pesquisas e contribui para uma visão mais ampla sobre o assunto *linked data*. Além disso, fornece vários insumos para enriquecer a discussão sobre o rumo das pesquisas e as prováveis tendências nesse campo de pesquisa.

2.3.3 Q3: Aplicações de Dados Governamentais

Como resposta à questão Q3, observa-se na Figura 2 que o tipo de resultado mais obtido utilizando-se dados governamentais foi procedimento ou técnica, seguido por relatório. Ferramenta ou notação e modelo descritivo ou qualitativo não obtiveram nenhuma pesquisa com uso de dados governamentais, o que indica lacunas a serem exploradas por pesquisadores ao relacionar o uso de dados governamentais com visualização de *linked data* e SIG.

ISBN 978-85-61115-09-8

No entanto, é importante reconhecer a limitação da pesquisa, no que diz respeito às palavras-chave, que na área de engenharia de software, não são padronizadas, podendo ser específicas de um segmento de conhecimento ou idioma. Portanto, devido à nossa escolha em utilizar palavras-chave e strings de pesquisa, há um risco de que alguns estudos relevantes tenham sido omitidos. Para novas pesquisas neste campo, esta revisão sistemática da bibliografia é relevante, ao permitir a visualização do enquadramento de futuras pesquisas em relação aos trabalhos já realizados.

3 Conclusão

Esta pesquisa apresentou os resultados de um estudo bibliométrico sobre os termos "visualização de *linked data*" e "Sistema de Informação Geográfica". Foi desenvolvido com o propósito de disponibilizar aos pesquisadores e interessados nos temas um mapeamento sobre as características das pesquisas que tratam dos assuntos abordados.

O estudo utilizou técnicas de revisão sistematizada da literatura para a captura de dados que, contextualizados, possibilitaram a identificação de padrões e tendências da literatura científica e mostrou, assim, que pesquisas deste tipo podem ser promissoras por auxiliarem os pesquisadores a identificar embasamento teórico nessa área de estudo. Essa técnica permite lidar com o desafio de agrupar informações e traçar perfis representativos no campo de estudo de *linked data* e SIG, além do enfoque da pesquisa no decorrer do tempo. Apesar de emergente, conclui-se que o número de pesquisas nesta área é crescente e se concentra principalmente no desenvolvimento de soluções como exemplos, mas ainda com pouco uso de dados governamentais. Esses pontos servem de insumo para os pesquisadores desta área nos próximos anos.

Assim, a partir de dados bibliográficos, o mapeamento das pesquisas sobre SIG e visualização de *linked data*, apresentado neste trabalho, promove uma maior compreensão sobre o histórico e o estado da arte nesta área de pesquisa, a nível internacional.

Referências

AHLBERG, C., WILLIAMSON, C., and SHNEIDERMAN, B. **Dynamic queries for information exploration: an implementation and evaluation**. pages 619-626, 1992.

CARD, S. K., MACKINLAY, J. D., and SCHEIDERMAN, B. **Readings in information visualization: using vision to think**. Interactive Technologies Series. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

CHEN, C., H¨ARDLE, W., and UNWIN, A. **Handbook of data visualization**. Springer Handbooks of Computational Statistics. Springer London, Limited, 2007.

COOK, D., MULROW, C., and HAYNES, B. **Systematic Reviews: Synthesis of Best Evidence for Clinical Decisions**, 1997.

GEROIMENKO, V. and CHEN, C. **Visualizing the Semantic Web: Xml-Based Internet and Information Visualization**. Springer-Verlag GmbH, 2003.

GLASS, G. V. **Primary, Secondary, and Meta-Analysis of Research**. Educational

ISBN 978-85-61115-09-8

Researcher, 5(10):3-8, 1976.

KING, W. and HE, J. **Understanding the role and methods of meta-analysis in is research**. Communications of the Association for Information Systems, 16(1):665-686, 2005.

KITCHENHAM, B., DYBA, T., and JORGENSEN, M. **Evidence-based software engineering**. Proceedings of ICSE 2004.

Apêndice

Quadro 2: Lista das Publicações

#	Publicação
1	A data model and query language for an extension of RDF with time and space. Manolis Koubarakis, Kostis Kyzirakos, Babis Nikolaou, Michael Sioutis, Stavros Vassos
2	A RESTful Proxy and Data Model for Linked Sensor Data. Krzysztof Janowicz, Arne Br¨oring, Christoph Stasch, Sven Schade, Thomas Everding, and Alejandro Llaves
3	Aggregating Geoprocessing Services using the OAI-ORE Data Model. Carlos Abargues, Carlos Granell, Laura Diaz, Joaquin Huerta
4	An Environment for the Conceptual Harmonisation of Geospatial Schemas and Data. Thorsten Reitz, Simon Templer
5	An implementation of a temporal and spatial extension of RDF and SPARQL. George Garbis, Konstantina Mpereta, Manos Karpathiotakis, Kostis Kyzirakos, Babis Nikolaou, Michael Sioutis, Stavros Vassos, Iris Miliaraki, Katerina Papadaki, Manolis Koubarakis
6	Applying Linked Data Technologies to Greek Open Government Data - A Case Study. Eleni Galiotoua, Pavlina Fragkou
7	Applying Semantic Linkage in the Geospatial Web. Aneta Florczyk, Francisco FLopez-Pellicer, Rub´en B´ejar, Javier Nogueras-Iso, Javier Zarazaga-Soria
8	Bridging the gap between user generated spatial content and the semantic web. Gianfranco Gliozzo
9	Data Models and Query Languages for Linked Geospatial Data. Manolis Koubarakis, Manos Karpathiotakis, Kostis Kyzirakos, Charalampos Nikolaou, and Michael Sioutis
10	Design and Development of a Mineral Exploration Ontology. Hilal Mentes
11	Discovery and Construction of Authors Profile from Linked Data. Atif Latif, Muhammad Tanvir Afzal, Denis Helic, Klaus Tochtermann, Hermann Maurer
12	Exploiting Semantics of Web Services for Geospatial Data Fusion. Pedro Szekely, Craig Knoblock, Shubham Gupta, Mohsen Taheriyan, Bo Wu
13	Explorative user interfaces for browsing historical maps on the Web. Rainer Simon, Joachim Korb, Christian Sadilek, Matthias Baldauf
14	Exploring the Research Field of GIScience with Linked Data. Carsten Keßler, Krzysztof Janowicz, and Tomi Kauppinen
15	GeoLinked Data and INSPIRE through an Application Case. Luis M. Vilches-Bl´azquez, Boris Villaz´on-Terrazas, Victor Saquicela, Alexander de Le´on, Oscar Corcho, Asunci´on G´omez-P´erez
16	GIS Supporting Data Gathering and Fast Decision Making in Emergencies Situations. Tiago Marino, Bruno Nascimento, Marcos Borges
17	Improving the Usability of Integrated Applications by Using Interactive Visualizations of Linked Data. Heiko Paulheim
18	Linked Environment Data. Thomas Bandholtz, Joachim Fock
19	Linked Open Data Aggregation - Conflict Resolution and Aggregate Quality. Tomas Knap, Jan Michelfeit, Martin Necasky
20	Linked open data in sensor data mashups. Danh Le-Phuoc, Manfred Hauswirth
21	Linked Open Piracy - A story about e-Science, Linked Data, and Statistics. Willem Robert van Hage, Marieke van Erp, V´eronique Malais´e
22	Linked Open Science - Communicating, Sharing and Evaluating. Tomi Kauppinen, Giovana Mira de Espindola
23	Linked Open Social Signals. Pablo N. Mendes, Alexandre Passanty, Pavan Kapanipathi, Amit P. Sheth
24	Online Dispute Resolution for the Next Web Decade – The Ontomedia Approach. Marta Poblet, Pompeu Casanovas, Jos´e Cobo
25	Open data in Finland: Public sector perspectives on open data. Paulina Lehtonen

ISBN 978-85-61115-09-8

26 Open Government Implementation Model. Bernhard Krabina, Thomas Prorok, Brigitte Lutz
27 Producing and Using Linked Open Government Data in the TWC LOGD. Timothy Lebo, John S. Erickson, Li Ding, Alvaro Graves, Gregory Todd Williams, Dominic DiFranzo, Xian Li, James Michaelis, Jin Guang Zheng, Johanna Flores, Zhenning Shangguan, Deborah L. McGuinness, Jim Hendler
28 Publishing and Interacting with Linked Data. Roberto Garc´ıa, Josep Brunetti Antonio L´opez-Muz´as, Manuel Gimeno, Rosa Gil
29 Reconstructing Semantics of Scientific Models - a Case Study Martine de Vos, Willem Robert van Hage, Jan Ros, Guus Schreiber
30 Semantic Aspects of EarthCube. Pascal Hitzler, Krzysztof Janowicz, Gary Berg-Cross, Leo Obrst, Amit Sheth, Tim Finin, Isabel Cruz
31 Semantic Web Portal - A Platform for Better Browsing and Visualizing Semantic Data. Ying Ding, Yuyin Sun, Bin Chen, Katy Borner, Li Ding, David Wild, Melanie Wu, Dominic DiFranzo, Alvaro Graves Fuenzalida, Daifeng Li, Stasa Milojevic, ShanShan Chen, M. Sankaranarayanan, Ioan Toma
32 Semantic Web Portal in University Research Community Framework. Rahmat Hidayat, Yazrina Yahya, Shahrul Azman Mohd Noah, Mohd Zakree Ahmad, Abdul Razak Hamdan
33 SemUNIT - French UNT and Linked Data.
Yoann Isaac, Yolaine Bourda, Monique Grandbastien
34 Sensor Discovery on Linked Data.
Joshua Pschorr, Cory Henson, Harshal Patni, Amit Sheth
35 Sharing geospatial provenance in a service-oriented environment. Peng Yue, Yaxing Wei, Liping Di, Lianlian He, Jianya Gong, Liangpei Zhang
36 Spatio-Temporal Aggregation of European Air QualityObservations in the Sensor Web. Christoph Stasch, Theodor Foerster, Christian Autermann, Edzer Pebesma
37 Sustainability through Open Data - Examples from Switzerland. Antoine Logean, Oleg Lavrovsky, Peter Gassner, Ralph Straumann
38 TELEIOS - A Database-Powered Virtual Earth Observatory. M. Koubarakis, K. Kyzirakos, M. Karpathiotakis, C. Nikolaou, S. Vassos, G. Garbis, M. Sioutis, K. Bereta, D. Michail, C. Kontoes, I. Papoutsis, T. Herekakis, S. Manegold, M. Kersten, M. Ivanova, H. Pirk, Y. Zhang, M. Datcu, G. Schwarz, O. C. Dumitru, D. E. Molina, K. Molch, U. D. Giammatteo, M. Sagona, S. Perelli, T. Reitz, E. Klien, R. Gregor
39 The Digital Earth as Knowledge Engine. Krzysztof Janowicz, Pascal Hitzler
40 The Information Workbench. Peter Haase, Andreas Eberhart, Sebastian Godelet, Tobias Mathab, Thanh Tran, Gunter Ladwig, Andreas Wagner
41 The Path is the Destination - Enabling a New Search Paradigm with Linked Data.
Jorg Waitelonis, Magnus Knuth, Lina Wolf, Johannes Hercher, Harald Sack
42 Towards a Research Framework - Using the Semantic Web for (In) Validating this Famous Geo Assertion. Stefan Hahmann, Dirk Burghardt, Beatrix Weber
43 Trentino Government Linked Open Geodata - A Case Study. Pavel Shvaiko, Feroz Farazi, Vincenzo Maltese, Alexander Ivanyukovich, Veronica Rizzi, Daniela Ferrari, Giuliana Ucelli
44 Ubiquitous Interaction And Collaboration With Touristic Services. Dieter Fensel, Christoph Fuchs, Jennifer Kaiser, Iker Larizgoitia, Birgit Leiter, Alex Oberhauser, Corneliu Stanciu, Ioannis Stavrakantonakis, Ioan Toma
45 Unifying Stream Data And Linked Open Data. Danh Le-Phuoc, Josiane Parreira, Michael Hausenblas, Manfred Hauswi
46 Upper Tag Ontology (UTO) For Integrating Social Tagging Data. Ying Ding, Elin Jacob, Michael Fried, Ioan Toma, Erjia Yan, Schubert Foo
47 Urban Mash-Ups. Daniele Dell'Aglio, Irene Celino, Emanuele Della Valle
48 Utilizing Open Content For Higher-Layered Rich Client Applications. Monika Steinberg, J¨urgen Brehm
49 Visinav Visual Web Data Search And Navigation. Andreas Harth
50 Design And Implementation Of A Gazetteer. Andr´e Soares
51 GI Systems For Public Health With An Ontology Based Approach. Nurefsan G¨ur
52 Providing Energy Efficiency Location-Based Strategies For Buildings Using Linked Open Data. Ana Sanchis Huertas
53 Representing Historical Knowledge In GIS. Karl Grossner
54 Geographic Feature Pipes. Marcell Roth
55 Modelos De Base De Datos De Grafo Y RDF. Renzo Angles Rojas

ISBN 978-85-61115-09-8

FABRICO/CIÊNCIA NO DESENVOLVIMENTO DE AMBIENTES LINKED DATA PARA A CIÊNCIA

Rafael Port da Rocha
rafael.rocha@ufrgs.br

Resumo
Fabrico/Ciência é um ambiente que vem sendo desenvolvido para explorar o uso de recursos da Web 2.0, da Web Semântica, de Linked Data e do modelo de interoperabilidade dos arquivos abertos (OAI-PMH) para promover o desenvolvimento da Ciência. Este artigo identifica experiências, ontologias e questões de pesquisa relacionadas com a publicação de dados e a descrição de recursos da Ciência em ambientes Linked Data; e apresenta Fabrico/Ciência e suas contribuições neste contexto. Destaca que funcionalidades de Fabrico/Ciência baseadas na Web 2.0, como fóruns, wikis e anotações semânticas, apoiam a construção coletiva de ontologias para a Ciência, bem como a coleta e descrição de recursos científicos; e que características temporais e de identificação das sentenças contribuem para a manutenção do histórico, para o registro e para identificação das mudanças nas ontologias, dados e descrições, e para assegurar a confiabilidade destes.

Palavras-chave: Linked Data. Linked Science. Ferramentas Linked Data.

Abstract
Fabrico/Ciência is an experiment that has being developed to explore the use of features of Web 2.0, Semantic Web, Linked Data, and OAI-PMH open archives interoperability model to promote the development of the Science. This article identifies experiments, ontologies and research issues related to the publication/description of science data and resources as linked data; and presents Fabrico/Ciência and its contributions in this context.It highlightsthat Web 2.0 based features of Fabrico/Ciência, as forums, wikis and semantic annotators, may support the collective construction of ontologies and collaborative actions to collect and describes scientific resources; and that temporal and identification features of sentences contribute to maintain the history of sentences, and to record and identify changes in the ontology, data and descriptions, and to support their reliability.

Key Words: Linked Data. Linked Science. Linked Data Tools.

Introdução

A web surgiu como uma rede global para publicação e interligação de documentos. Hoje, está evoluindo para um ambiente em que seus usuários passam também a ser seus construtores (web 2.0), em que máquinas (agentes de softwares inteligentes) passam a poder compreender o significado de seus recursos (Web Semântica), e em que dados, e não somente documentos, passam a ser interligados e compreensíveis por máquinas (Linked Data).

A Web Semântica e Linked Data trazem novas possibilidades para a comunicação e o intercâmbio de recursos da ciência. Enquanto que federações de bases de dados científicasestão focadas na publicação e na interoperabilidade de recursos bibliográficos, a

ISBN 978-85-61115-09-8

plataforma Web Semântica/Linked Data possibilita a interoperabilidade semântica que engloba também outros tipos recursos, como dados de pesquisa, pesquisadores, instituições, projetos, grupos de pesquisa, procedimentos, instituições de apoio, fomentos, eventos, cursos, etc.

Vivo (DEVARE,2007) e Eagle-I (TORNIAI, 2011) são exemplos de ambientes para Web Semântica em que recursos da ciência de várias universidades são semanticamente descritos, formando grandes redes de recursos interlidados. Esses ambientes ampliam as possibilidades de trocas no desenvolvimento da ciência, promovem a descoberta de recursos e aproximam pesquisas em um cenário distribuído. Segundo Colon e Holmes (2012), na ciência moderna, como nunca, uma equipe de alto nível requer maximizar o processo de descoberta:

> A formação de uma equipe de nível mundial requer de recursos ricos e varia dos para descoberta no campo da pesquisa, através dos quais seus cientistas membros possam ter uma consciência ampla e profunda do que está acontece ndo em sua disciplinas. Estão encerrados os tempos de simplesmente manter -se atualizado com a literatura atual, ou mesmo antecipar-se sobre o que está vindo ao ser editor de uma revista[...]. O problema com essa abordagem é qu e ela só pode fornecer um ponto de vista vantajoso a partir de uma pessoa, de uma rede ou de um quadro de referência, e não favorece a descoberta de proj etos para além do seu horizonte imediato (tradução nossa). (Colon e Holmes, 2012)

Fabrico/Ciência[25] é um experimento que está sendo desenvolvido que explora o uso de recursos da Web 2.0, da Web Semântica, de Linked Data, e de interoperabilidade entre bases de dados documentais (OAI-PMH), na promoção da ciência. Este artigo apresenta o ambiente Fabrico/Ciência, identificando seus recursos e funcionalidades para o desenvolvimento de ambientes para descrever recursos da ciência em Linked-Data, no que diz respeito à produção de ontologias a serem usadas para descrever os recursos da ciência, à anotação (produção) de recursos da ciência e à descrição da proveniência dos mesmos (visando garantir sua autenticidade e confiabilidade).

O artigo primeiramente (seção 1) apresenta um cenário envolvendo recursos da ciência em Linked Data, destacando ontologias para descrever e interligar recursos da ciência (incluindo recursos bibliográficos); esforços para publicar recursos bibliográficos como Linked Data; ambientes em que instituições descrevem e interligam diversos tipos de recursos da ciência; o desenvolvimento de ontologias para esses ambientes; e a coleta e a produção de descrição de recursos da ciência. Com base nesse cenário, identifica desafios relacionados ao desenvolvimento ambientes para descrição de recursos da ciência, enfatizando demandas por ferramentas e funcionalidades requeridas para esses ambientes. A seção 2 apresenta a arquitetura do ambiente Fabrico/Ciência, e a seção 3 analisa recursos de Fabrico/Ciência para atender o desenvolvimento de ontologias para ciência, a produção, importação e curadoria de descrições relacionadas a Ciência, e a garantia da confiabilidade e autenticidade das descrições via mecanismos de proveniência, relacionando o com características de outros ambientes.

1 Linked Data e a Ciência

Várias Ontologias e ambientes estão sendo desenvolvidos para descrever e interligar recursos da ciência na forma Linked Data. Com relação aos recursos bibliográficos, bibliotecas e bases de dados experimentam publicar seus registros em RDF. Por exemplo, a Biblioteca Britânica (DELIOT, 2014) explora a publicação em RDF de uma parte de seu

[25] Fabrico/Ciência – http://www.ufrgs.br/fabrico/ciencia

ISBN 978-85-61115-09-8

catálogo bibliográfico (aprox.: 2 milhões de registros), e o projeto VIAF (BOURBON, 2013) é uma iniciativa em que várias bibliotecas nacionais experimentam a publicação de seus registros de autoridade como Linked Data. O Projeto Europeana (HASLHOFER, 2011), que desenvolve uma federação de repositórios de bens culturaisda Europa, construída via protocolo OAI-PMH, experimenta a publicação em RDF dos metadados colhidos nos seus repositórios membros. Ensaios também são realizados para a publicação de registros de bases de dados referenciais, como ocorre com a baseJISC Open Citations (SHOTTON, 2013). .

BiBO[26] e CiTO[27] são ontologias usadas, respectivamente, para descrever tipos de recursos bibliográficos (artigos, teses, etc.) e tipos de relações de citações. A ontologia FaBiO[28] está baseada nos requisitos funcionais para descrição de registros bibliográficos, na qual um item é visto a partir de sua obra, de como essa obra nele é expressada e em como essa expressão nele se manifesta. SPAR[29] é uma ontologia que busca reunir de ontologias para publicação, envolvendo BiBO, CiTO e FaBIO, entre outras, sendo utilizada no projeto JISC Open Citations.

Projetos têm sido desenvolvidos, em âmbito institucional ou na forma de redes interinstitucionais, para descrever e interligar recursos da ciência, visando atroca e a descoberta de recursos naprodução da ciência, aproximando pesquisas e pesquisadores. Projetos com essa características descrevempesquisadores, projetos de pesquisa, órgãos, entidades, laboratórios, orientações, eventos, cursos, etc. LODUM (KAUPPINEN, T.; BAGLATZI, A; KEßLER, C., 2012) é um exemplo de projeto institucional, em que os recursos da ciência da universidade de Muenster são descritos, interligando os órgãos da universidade, pessoas (como pesquisadores) e produtos (como publicações). VIVO (DEVARE,2007) e Eagle-i (TORNIAI, 2011) são exemplos de ambientes de descrição de recursos da ciência em rede, envolvendo universidades e laboratórios. Eagle-I é uma plataforma de descoberta para auxiliar cientistas biomédicos a buscar e encontrar recursos previamente invisíveis, mas altamente valiosos. VIVO é uma rede que viabiliza a colaboração e a descoberta, envolvendo cientistas de várias disciplinas, descrevendo e interligando recursos da ciência de várias instituições de pesquisa.

Para descrever e interligar recursos da ciênciadestacam-se as ontologias dos ambientes VIVO e Eagle-I, e a derivada do modelo CERIF[30]. A ontologia ERO (TORNIAI, 2011), de Eagle-I, tem como foco principal representar recursos comumente usados na ciência, mas raramente compartilhados, incluindo reagentes, protocolos, instrumentos, especialidades, software, oportunidades de treinamento, estudos e espécies biológicas.

O desenvolvimento de ERO envolveu uma diversidade de conceitos e o reuso de várias ontologias. Para facilitar o processo de modelagem e o reuso de ontologias existentes, a elaboração de ERO envolveu o uso da ontologia de alto nível BFO[31] e deuma técnica de modularização de ontologia, com a especificação de um núcleo da área biomédica(TORNIAI, 2011). A técnicaMIREOT (COURTOT, 2011) foi usada para a importação de partes de ontologias reutilizadas, para evitar importações de classes e propriedades desnecessárias. (TORNIAI, 2011).

A ontologia de VIVO (CONLON e CORSON-RIKERT, 2012)está voltada à integração e ao compartilhamento de informação entre pesquisadores, envolvendo a descrição de dados, conceitos, pesquisadores, projetos, estudos, financiamentos, equipamentos,

[26] BiBO - Especificação em http://bibliontology.com/
[27] CiTO- Citation Typing Ontology – Especificação em http://purl.org/spar/cito
[28] FaBiO - Especificação em http://purl.org/spar/fabio
[29] SPAR - Semantic Publishing and Referencing Ontologies. http://sempublishing.sourceforge.net/#1
[30] CERIF - Common European Research Information Format: http://www.eurocris.org/ontologies/cerif/1.3
[31] BFO - Basic Formal Ontology – Especificação em http://www.ifomis.org/bfo/

ISBN 978-85-61115-09-8

serviços, artigos, trabalhos acadêmicos, etc. Seu desenvolvimento não utiliza ontologia de alto nível, havendo a reutilização de ontologias gerais, como FOAF[32] (para descrever pessoas/entidades), Event[33] (para eventos), SKOS[34] (para conceitos), e da ontologia de recursos bibliográficos BiBO. Uma junção entre VIVO e Eagle-I está sendo desenvolvida através da ontologia VIVO-ISF(TORNIAI, 2013). Essa junção exigiu uma estratégia baseada na refatoração (TORNIAI, 2013), com entidades relevantes sendo agrupadas em módulos básicos, cujo processo é apoiado por uma funcionalidade adicional, desenvolvida para a ferramenta Protegè, que permite anotações sobre os elementos da ontologia em construção e decisões de refatoração, pela equipe de desenvolvimento.

CERIF é um modelo de descrever informações de pesquisa (incluindo publicações, projetos, organizações, pessoas, produtos, patentes, serviços e eventos), recomendado pela Comunidade Europeia, cuja expectativa é prover uma grande quantidade de informação para conduzir a ciência em âmbito local, nacional e internacional. CERIF Ontology é uma ontologia baseada no modelo CERIF, e esforços estão sendo realizado para estabelecer mapeamentos entre esta ontologia e VIVO (LEZCANO et al, 2013).

O termo Linked Science tem sido usado para designar uma abordagem onde recursos científicos (fluxos de trabalho, processos, modelos, dados, métodos e métricas de avaliação) são semanticamente anotados e interconectados (KAUPPINEN, BAGLATZI, KEßLER, 2013). Elementos de pesquisa, seus contextos e interconexões (como publicação, pesquisa, pesquisador, hipótese, conclusão e dado) são especificados através da ontologia LSC[35]. Essa ontologia permite descrever os conteúdos de um artigo de pesquisa, relacionando pesquisas, hipóteses, previsões, experimentos, dados e publicações. A publicação de dados da ciência proporciona a sua reutilização, assim como a reprodução de experiências e a revalidação de seus resultados.

A proveniência é outro tema relevante na publicação e descrição de recursos da ciência. Proveniência sobre um recurso compreende em informações sobre suas origens, com indagações sobre "quem", "como", "o que" e "quando", envolvendo entidades, atividades e agentes que participam da sua produção. Essas informações são vitais para a garantia da qualidade e da confiabilidade do recurso. A ontologia PROV-O[36] é proposta pelo W3C para representar informações de proveniência em Linked Data.

Em Linked Data, informações de proveniência envolvem não somente recursos da web, mas também dados e conjuntos de dados representados através de sentenças RDF. Para possibilitar a descrição de sentenças e conjuntos de sentenças, estes necessitam ser identificados como recursos. Os recursos existentes em RDF para identificar sentenças e conjuntos de sentenças são a reificação (em que uma sentença é considerada um recurso) e os grafos nomeados (grafos RDF que são atribuídos a nomes, na forma de URI).Entretanto há uma indefinição quanto à presença desses instrumentos ou de outros instrumentos de identificação na próxima versão de RDF, e isso traz problemas para o estabelecimento e consolidação de princípios e recomendações referentes à proveniência (ECKERT ; PFEFFER e VÖLKER, 2010). ECKERT (2013) introduz o conceito de contexto de proveniência, aplicado a um conjunto de triplas que compartilham mesma proveniência.

Outra questão importante é a coleta e a curadoria de dados da ciência. Por exemplo, o projeto Eagle-I necessita coletar, gerenciar e organizar dados de pesquisa como organismos, vírus, reagentes, estudos, serviços, protocolos, de forma colaborativa e interinstitucional. Para

[32]FOAF – Especificação em http://xmlns.com/foaf/spec
[33]Event Ontology – Especificação em http://motools.sourceforge.net/event/event.htm
[34] SKOS - Simple Knowledge Organization System – Especificação em http://www.w3.org/TR/skos-reference
[35] LSC - Linked Science Core – Especificação em http://linkedscience.org/lsc/ns/
[36] PROV-O - ww.w3.org/TR/prov-o/

ISBN 978-85-61115-09-8

tal, um ambiente dirigido por ontologia foi desenvolvido, juntamente com a elaboração de diretrizes e fluxos de trabalho (VASILEVSKY,2012).

A partir do cenário apresentado, várias oportunidades e desafios são identificados, relacionados ao desenvolvimento de ambientes para descrição de recursos da ciência, enfatizando demandas por ferramentas e funcionalidades requeridas para esses ambientes. São eles:

- A existência de uma variedade de ontologias a serem usadas na descrição de recursos da ciência demanda por técnicas e ferramentas para alinhamento, mapeamento, fatoração, modularização e importação de ontologias. Demanda também pelo uso de ontologias de alto nível, por ferramentas para anotar ações e decisões do processo de construção colaborativa da ontologia,a e para explorar e anotar ontologias a serem reutilizadas.
- A necessidade por ambientes interinstitucionais para descrever e interligar recursos da ciência, como VIVO e Eagle-i, demanda por esforços para estabelecer processos para coleta, anotação e curadoria de dados e recursos da ciência, envolvendo ambientes colaborativos interinstitucionais, e com apoio de ferramentas e de técnicas.
- A existência de registros bibliográficos armazenados em repositórios que suportam serviços para interoperabilidade de metadados, como o protocolo de colheita de metadados OAI_PMH, demanda por ambientes para descrição de recursos da ciência com mecanismos para importar e disponibilizar esses registros em RDF, via protocolos de interoperabilidade de metadados. Além disso, a existência de esforços de bibliotecas nacionais e de bases de dados bibliográficas para publicar dados bibliográficos como Linked Data leva a necessidades por ambientes para descrição de recursos da ciência com funcionalidades para estabelecer interligações com esses dados bibliográficos.
- A coleta, importação e descrição de dados e recursos demandam por mecanismos para assegurar autenticidade, credibilidade e qualidade às informações, envolvendo ontologias, princípios e recomendações de proveniência, assim como recursos para identificar sentenças RDF (para permitir a descrição de sua proveniência).

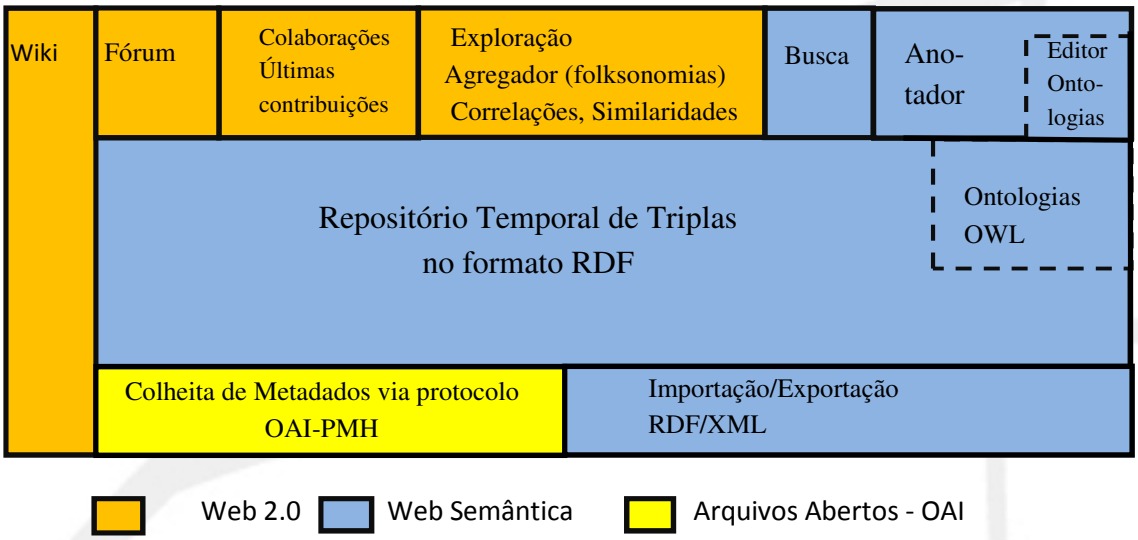
2 Fabrico/Ciência

Fabrico/Ciência é uma plataforma para descrição de recursos da ciência em Linked Data. Decorre da necessidade por um ambiente flexível, focado na construção coletiva, que envolve o desenvolvimento/reuso de ontologias da ciência, a anotação semântica de recursos da ciência e a colheita de registros bibliográficos armazenados em repositórios. Seu repositório registra a autoria e o histórico das anotações armazenadas.

Fabrico/Ciência (figura 1) é um anotador baseado em ontologias construído de acordo com a arquitetura da Web Semântica, adotando as linguagens RDF, RDFS e OWL, e seguindo o modelo de negócio de sistemas para Web 2.0 de O'REILLY (2007), sendo um serviço para web que explora a arquitetura de participação e a inteligência coletiva. O anotador gera dinamicamente e incrementalmente os formulários da interface de anotação, a cada momento em que o usuário passa a descrever o recurso. Por exemplo, quando um usuário descreve que um recurso é instância de uma determinada classe, a interface de anotação passa a incluir campos de anotação referentes às características e às restrições especificadas na classe. O ambiente de desenvolvimento de ontologias de Fabrico/Ciência é obtido através da configuração do anotador, isto é, o anotador é configurado para gerar formulários de anotação de classes e propriedades, a partir das ontologias que definem as linguagens RDFS e OWL (ROCHA, 2010).

ISBN 978-85-61115-09-8

Figura 1 – Arquitetura de Fabrico/Ciência



Fonte: autor

Fabrico/Ciência utiliza técnicas da Web 2.0 para capturar contribuições dos usuários (via anotações semânticas) e para auxiliar a exploração dos dados. Incorpora um ambiente wiki, um fórum de discussão, funcionalidades de agregação típicas de folksonomias, e um mecanismo que informa as últimas contribuições e que relaciona usuários que realizaram contribuições envolvendo mesmos recursos. Forums e wikis podem ser criados para discutir e documentar qualquer recurso descrito, inclusive sentenças (via reificação), ou classes (no ambiente de construção de ontologias).As mensagens dos fóruns são sentenças produzidas de acordo uma ontologia que segue o modelo Issue Based Information System, de Grant (1982).

Para facilitar a exploração das informações, o ambiente possui funcionalidades de agregação (figura 2) que estão baseadas nas técnicas utilizadas na Web 2.0 para analisar folksonomias. Essas agregações permitem que valores de propriedades possam ser apresentados em nuvens (em que o tamanho da letra corresponde à sua ocorrência), assim como a exibição de recursos que se relacionam por serem objetos de descrições que envolvem mesmos sujeitos, como cotermos e coautoria (assuntos ou autores que são relacionados por terem sido atribuídos a mesmos recursos). Também permite identificar relações não explícitas entre recursos através de cálculos de similaridades, envolvendo recursos com valores similares de propriedades (como, a partir de um autor ou assunto, identificar autores e assuntos similares).

ISBN 978-85-61115-09-8

Figura 2 – Agregações em Fabrico/Ciência. Recursos Bibliográficos colhidos da Revista EmQuestão

Fonte: autor

Fabrico/Ciência possibilita a colheita de metadados de repositórios via do protocolo OAI-PMH, assim como o apoio à preparação dos dados colhidos (ROCHA, 2013 e ROCHA 2012). Por exemplo, permite a definição de regras que usam expressões regulares para normalizar valores colhidos, como, por exemplo, transformar uma notação de nome de "Sobrenome, Nome" para "Nome Sobrenome".

O repositório de Fabrico/Ciência armazena sentenças uma versão estendida da classe rdf:Sentence.Uma sentença no Fabrico/Ciência é composta pelas propriedades sujeito (rdf:subject), predicado (rdf:predicate) e objeto (rdf:object), adicionadas de identificador da sentença, modelo (conjunto em que a sentença foi criada), criador/atualizador da sentença, versão da sentença e período de vigência da versão.

Essa estratégia de adotar um repositório temporal de sentenças, com retificação e identificação de conjuntos de sentenças, trouxe grandes benefícios e funcionalidades ao sistema. Com isso, Fabrico/Ciência dispõe de instrumentos para garantir a confiabilidade e a autenticidade das sentenças, visto que sempre identifica o autor das mesmas e suas alterações. Também possibilita o rastreamento das alterações e o acompanhamento da evolução histórica do que está sendo construído (incluindo a ontologia). Viabiliza também um ambiente de construção coletivaque registra as mudanças, e possibilita que sentenças sejam sujeitas a anotações, recebendo comentários, avaliações e indicações de estado, etc.

No repositório do Fabrico/Ciência, uma sentença é armazenada em uma tabela de banco de dados relacional. Um desenvolvimento preliminar buscou a utilização do repositório triplas RDF Jena, desenvolvido por MCBRIDE (2002). Entretanto essa estratégia foi descontinuada devido à complexidade e baixo desempenho do mecanismo de reificação, e para permitir que o sistema possa operar em uma configuração passada da base de dados.

ISBN 978-85-61115-09-8

3 Fabrico/Ciência em Ambientes Linked Data para Ciência

Esta seção relaciona Fabrico/Ciência a questões referentes ao desenvolvimento de ambientes para descrição de recursos da ciência, enfatizando demandas por ferramentas e funcionalidades requeridas para esses ambientes no que dizem respeito: ao desenvolvimento de ontologias para ciência em um ambiente colaborativo; à produção, importação e curadoria de descrições relacionadas a Ciência; e à garantia da confiabilidade e autenticidade das descrições via mecanismos de proveniência.

O desenvolvimento das ontologias de ERO (Eagle-I) e VIVO-ISF, em função da quantidade e da diversidade de ontologias envolvidas, exigiu a especificação de técnicas e procedimentos específicos para importação, modularização e refatoração, com funcionalidades de algumas dessas técnicas sendo incorporadas na ferramenta de desenvolvimento da ontologia. Na construção de ERO, a ferramenta Protegè foi estendida para incorporar uma técnica própria para importar ontologias (TORNIAI, 2011). Essa mesma ferramenta foi estendida para gerenciar a refatoração, no processo de desenvolvimento de VIVO-ISF (TORNIAI, 2013). Fabrico/Ciência é um ambiente para a construção de ontologias que permite o desenvolvimento de extensões para incorporar novos métodos associados ao processo. Isso pode ser feito através da especificação de uma ontologia que representa o método, e da utilização desta ontologia para configurar o anotador, para que este passe a gerar formulários de anotação referentes ao método.

Fabrico/Ciência também favorece a construção coletiva de ontologias, pois dispõe de ferramentas colaborativas, como wikis e fóruns,que podem ser associados a elementos da ontologia em desenvolvimento. Também dispõe de funcionalidades que mantêm os usuários informados sobre as ultimas contribuições e que relacionam usuários com base nas suas contribuições (aproximando usuários). Por ter características temporais, o ambiente permite o acesso a todos os estados da ontologia e seus elementos, isto é, ao histórico e às mudanças na ontologia.

Fabrico/Ciência oferece vantagens para a coleta, anotação e curadoria de descrições, pois seus recursos de colaboração, de configuração do anotador, e de armazenamento temporal de sentenças permitem, respectivamente, o incentivo à construção coletiva, o desenvolvimento de extensões que incorporem métodos e técnicas de importação, coleta e curadoria, e o acesso ao histórico e às mudanças ocorridas.

Fabrico/Ciência também oferece recursos para gerenciar a confiabilidade e a autenticidade das descrições armazenadas. Devido às suas características temporais, todos os estados da base de dados são mantidos, permitindo a rastreabilidade frente às atualizações realizadas nas informações. Devido à representação estendida da sentença, que identifica também a sentença, seu autor, e o conjunto em que foi esta criada;descriçõesde proveniência podem ser atribuídas tanto à sentença quanto a conjuntos de sentenças que possuem origens comuns. Ontologias de proveniência podem ser desenvolvidas/utilizadas para documentar informações de proveniência.

Quanto ao desenvolvimento coletivo de ontologias, características de Fabrico/Ciência estão presentesno ambiente Collaborative Protejè,(TUDORACHE; NOY; MUSEN, 2008), que permite o desenvolvimento colaborativo de ontologias através de fóruns de discussões eda anotação de componentes da ontologia e de mudanças.Collaborative Protejè, entretanto, é específico para o desenvolvimento de ontologias, enquanto que Fabrico/Ciência permite também a anotação, coleta e curadoria, de forma colaborativa.

Vitro é a plataforma para desenvolvimento utilizada pelo ambiente VIVO, que integra três funções em umaúnica ferramenta: criação, importação e edição de ontologias;importação ou criação/edição interativa de sentenças RDF, emconformidade com a ontologia; e exibição do conteúdo na Web,suportando funções de navegação e consulta(LOWE, 2011). Ao contrário de Fabrico/Ciência, o seu editor de ontologias não explorarecursos de construção

ISBN 978-85-61115-09-8

colaborativa da Web 2.0 (wikis, foruns) e o registros de mudanças. Seu ponto forte é permitir a configuração de um ambiente para exibição do conteúdo na construção de um web site.

Ontowiki é um ambiente para Web Semântica usado para apresentação, autoria e gestão de bases de conhecimento, inspirado nas facilidades que os ambientes wiki oferecem para o trabalho colaborativo. Possui, como principais caraterísticas, oferecer mecanismos para a apresentação e a edição intuitiva dos dados;para gerar diferentes visões e agregações da base de conhecimento; para gerenciar (rastrear, rever e reverter seletivamente) mudanças; para promover o desenvolvimento colaborativo (envolvendodiscussões, votações);para prover estatísticas online; e para distribuir de informação (HEINO, 2009). Em OntoWiki, cada sentença pode ser anotada, comentada e avaliada por usuários. O acesso às sentenças por partes dos usuários é usado prover visões da base de dados baseadas na popularidade, e o sistema registra quem realizou as contribuições à ontologia, comentários ou adição de instâncias (AUER, S; DIETZOLD, S; RIECHERT, T, 2006). Os conteúdos da base de conhecimento são editados e navegados através de telas que são geradas automaticamente a partir da ontologia. Em Ontowiki, o versionamento da base de conhecimento é gerenciado no âmbito das sentenças RDF. Nesseambiente um recurso é usado para reunir todas as alterações que foram realizadas na sentença (informações de sentença adicionada, removida e sentença alterada).

OntoWiki apresenta funcionalidades que atendem ao desenvolvimento de ontologias para ciência em um ambiente colaborativo; a produção de descrições relacionadas a Ciência; e a garantia da confiabilidade e autenticidade das descrições. É similar ao Fabrico/Ciência ao permitir a anotação de sentença, uso de recursos como fóruns. A principal diferença entre os dois ambientes está na estrutura para registrar alterações. Fabrico/Ciência armazena todas as versões de uma sentença, permitindo a configuração da base de dados em recortes temporais, enquanto que Ontowiki registra todas as ações que determinaram criação, alterações e remoção de uma sentença.

Considerações Finais

A Descrição de recursos da Ciência é um campo de pesquisa promissor e desafiador. Este artigo apresentouFabrico/Ciência e suas contribuições neste contexto. Suas funcionalidades baseadas na Web 2.0 e seus recursos para anotações semânticasapoiam a construção coletiva de ontologias para a Ciência, bem como a coleta, a descrição e a curadoria de recursos e dados científicos. Características temporais e de identificação das sentenças contribuem para a manutenção do histórico, para o registro e a identificação das mudanças em ontologias, dados e descrições, e para assegurar a confiabilidade destes.

Um experimento foi realizado em Fabrico/Ciência com a importação, via OAI-PMH, de registros bibliográficos de uma Revista Eletrônica (ROCHA, 2012). No momento, estudos estão sendo desenvolvidos, com apoio do ambiente (coletando e anotando ontologias e documentos), para identificar e analisar ontologias para descrever recursos da Ciência. Os próximos passos serão usar o ambiente para a construção de uma ontologia para descrever recursos da Ciência. A seguir serão estabelecidas diretrizes e métodos para coletar, descrever, curar e garantir a confiabilidade de informações da ciência,para então realizar a coleta e descrição de recursos da Ciência ligados à área da Ciência da Informação

Referências

AUER, S; DIETZOLD, S; RIECHERT, T. OntoWiki–a tool for social, semantic collaboration. In: The Semantic Web-ISWC 2006. **Proceedings…** Springer Berlin Heidelberg, 2006.

ISBN 978-85-61115-09-8

BOURDON. F. VIAF: A hub for a multilingual access to varied collections. In World Library and Information Congress, 2011, San Juan, Puerto Rico, **Proceedings…** 2011

COLON, M; HOLMES, K. Big Science teams built on research discovery and networking systems. **The Academic Executive Brief**, v.2, n.2, 2012

CONLON, M; CORSON-RIKERT, J. VIVO: **A Semantic Approach to Scholarly Networking and Discovery**. Morgan & Claypool Publishers, 2012.

COURTOT, M. et al. MIREOT: The minimum information to reference an external ontology term. **Applied Ontology**, v. 6, n. 1, p. 23-33, 2011.

DELIOT C. **Publishing the British National Bibliography as Linked Open Data.** The British Library Collection Metadata, 2014

DEVARE, M.; et ali.VIVO: Connecting People, Creating a Virtual Life Sciences Community. **D-Lib Magazine**, v. 3, n. 7/8, 2007

ECKERT, Kai. Provenance and Annotations for Linked Data. In DCMI International Conference on Dublin Core and Metadata Applications, Lisboa, Portugal. **Proceedings…** 2013

ECKERT, Kai; PFEFFER, Magnus; VÖLKER, Johanna. Towards Interoperable Metadata Provenance. In: Proceedings of the ISWC workshop on Semantic Web for Provenance Management (SWPM). **Proceedings…** 2010.

GRANT, D. Issue-Based Information System (IBIS). In. OLSEN, S. **Group planning and problem-solving methods in engineering management**, New York: Wiley-Interscience, 1982. p. 203-246

HASLHOFER, Bernhard; ISAAC, Antoine. data. europeana. eu: The europeana linked open data pilot. In: International Conference on Dublin Core and Metadata Applications. **Proceedings…** 2011.

HEINO, N. et al. Developing semantic web applications with the ontowiki framework. In. **Networked Knowledge-Networked Media.** Springer Berlin Heidelberg, 2009.

KAUPPINEN, T.; BAGLATZI, A.; KEßLER. C. Linked Science: Interconnecting Scientific Assets. In CRITCHLOW, Terence; VAN DAM, Kerstin Kleese (Ed.). **Data-Intensive Science**. CRC Press, 2013.

LEZCANO, Leonardo et al. Promoting International Interoperability of Research Information Systems: VIVO and CERIF. **Journal of Universal Computer Science**, v. 19, n. 12, p. 1854-1867, 2013.

LOWE, Brian et al. The Vitro Integrated Ontology Editor and Semantic Web Application. In:International Conference on Biomedical Ontologies. **Proceedings…** 2011.

MCBRIDE, Brian. Jena: A semantic web toolkit. **IEEE Internet computing**, v. 6, n. 6, p. 55-59, 2002.

O'REILLY, T. What is Web 2.0: design patterns and business models for the next generation of software. **Journal of Digital Economics**, n. 65, 2007.

ROCHA, R . Desenvolvimento de Ontologias apoiado pela anotação semântica de textos. In. Seminário de Pesquisa em Ontologias no Brasil,3, Florianópolis, 2010. **Anais...** 2010.

ROCHA, R. FABRICO/CIÊNCIA: Um Ambiente Linked Data para o Mapeamento da Ciência. **Em Questão**. v.18, n. 3, 2012

ISBN 978-85-61115-09-8

ROCHA, R. Web semântica, dados ligados e web 2.0: explorando novas fronteiras para os arquivos abertos. In. Encontro Nacional de Pesquisa em Ciência da Informação, 14, Florianópolis, **Anais...**Florianópolis. 2013

SHOTTON D. Open citations. **Nature**, 502: 295–297., 2013

TORNIAI, Carlo et al. Developing an Application Ontology for Biomedical Resource Annotation and Retrieval: Challenges and Lessons Learned. In: International Conference on Biomedical Ontology, 2, Buffalo, NY, USA, 2011. **Proceedings…** CEUR-WS, v.833, 2011

TORNIAI, Carlo et al. Finding common ground: integrating the eagle-i and VIVO ontologies. In: International Conference on Biomedical Ontology, Montreal, Quebec, 2013. **Proceedings…** CEUR-WS, V. 1060, 2013.

TUDORACHE, Tania; NOY, Natalya Fridman; MUSEN, Mark A. Collaborative Protege: Enabling Community-based Authoring of Ontologies. In: International Semantic Web Conference. **Proceedings…** 2008.

VASILEVSKY, Nicole et al. Research resources: curating the new eagle-i discovery system. **Database**, v. 2012, p. bar067, 2012.

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

AS TECNOLOGIAS OPEN ARCHIVES INITIATIVE – OBJECT REUSE AND EXCHANGE E LINKED OPEN DATA: características ecomplementaridades

Joel de Souza
Joeldesouza@grad.ufsc.br

Resumo
Este estudo propõe uma análise das tecnologias *Open Archives Initiative - Object Reuse and Exchange* (OAI-ORE) e *Linked Open Data* (LOD), verificando características, similaridades e diferenças. A Iniciativa dos Arquivos Abertos é responsável pelo desenvolvimento de padrões de interoperabilidade culminando na disseminação de conteúdo dos repositórios digitais. O OAI-ORE, propende permitir que objetos complexos sejam reutilizados e trocados entre repositórios, transcendendo a ideia de apenas hospedar conteúdo estático. Enquanto o LOD tem como função primordial publicar e estruturar dados da Web. Os estudos e análises sobre estas tecnologias se justifica, em função da crescente demanda por informações específicas e particularizadas com seu consequente armazenamento em repositórios digitais visando sua posterior difusão, no fornecimento de subsídios à pesquisa científica. Os objetivos deste estudo compreendem estudar e descrever as tecnologias OAI-ORE e LOD por meio de levantamento, identificação e análise de documentos impressos e eletrônicos. A análise de conteúdo de Bardin foi empregada visando à obtenção de elementos para o desenvolvimento dos estudos a respeito destas tecnologias. Em campo acadêmico-científico, a reutilização e troca de objetos digitais mostra-se como mais uma alternativa que, proporciona otimização das potencialidades de recuperação e disseminação da informação. Por fim, concluiu-se que ambas as tecnologias podem se complementar organizando e reestruturando conteúdo da web.
Palavras-chave: LOD. OAI-ORE. Tecnologias da Informação.

Abstract
This studyproposesan analysis oftechnologiesOpenArchives Initiative-ObjectReuseand Exchange(OAI-ORE) and LinkedOpen Data(LOD), checking characteristics, similaritiesand differences. TheOpen ArchivesInitiativeisresponsible for developinginteroperability standardsculminatingin the disseminationof digitalcontentrepositories.The OAI-ORE, is inclined to allow complex objects to be reused and exchanged between repositories, transcending the idea of just staying static content. While the LOD has the primary structure data publishing and web function. Studiesandanalyzes ofthese technologiesis justified, given the increasingdemand forspecificandparticularizedinformationwith yoursubsequentstoragein digital repositoriesseekingits subsequentdiffusionin providingsubsidiesfor scientific research. The objectives of this study include study and describe the OAI-ORE LOD and technologies through survey, identification and analysis of printed and electronic documents. Content analysis of Bardin was employed in order to obtain elements for the development of studies about these technologies. In academic and scientific field, reuse and exchange of digital objects shows up as an alternative that provides optimization of the potential recovery and dissemination of information.Finally, it was concluded that both technologies can complement restructuring and organizing Web content.

Key Words: LOD. OAI-ORE. Information Technology.

1.Introdução

ISBN 978-85-61115-09-8

A proposta da OAI - *Open Archives Initiative* desde a sua concepção, emoutubro de 1999, pela Convenção de Santa Fé, realizada no Novo México/EUA, foia de viabilizar o acesso irrestrito a informação certificada e disponibilizadaincondicionalmente na rede mundial de computadores, como prevêem seusprincípios de: auto-publicação; armazenamento em longo prazo; política de gestãocom foco na preservação de objetos digitais; acesso livre e irrestrito aos metadados;uso e desenvolvimento de padrões e protocolos com vistas à promoção dainteroperabilidade e fomento ao uso de *software Open Source* - OS (VAN DESOMPEL; LAGOZE, 2000).

Com o projeto OAI-ORE, iniciado em outubro de 2006, é lançada aperspectiva de os repositórios digitais abandonarem o estigma de simples depósitosde informação para se projetarem ao posto de eficazes subsidiadores doconhecimento. Trata-se de uma iniciativa de colaboração internacional visando ointercâmbio de informações contidas nos objetos digitais compostos, constituindo-se este, um dos diferenciais em relação ao *Protocol Metadata Harvesting do Open Archives Initiative* (OAI-PMH), que tem foco exclusivo nos metadados (LAGOZE etal., 2008).

Com interesse centralizado em novas tecnologias da informação queproporcionem maior interação entre sistemas informáticos e pesquisadores, sãovisualizadas perspectivas de possíveis implementações que se justificam pelaimportância que possuem os repositórios digitais na produção científica. Aspectostais como o de preservação digital e intercâmbio de objetos complexos entresistemas, asseguram maior interoperabilidade e acesso longínquo aos documentose seu conteúdo informacional, conforme prevêem as especificações do projeto OAI-ORE,(*OPEN ARCHIVES INITIATIVE*, 2008).

Esta tecnologia propende integrar todas as formas emergentes de informação em rede e os recursos trabalhados estão abarcados nosformatos: PDF, DOC, HTML, JPEG, MP3, compreendendo arquivos de texto,imagem, áudio, vídeo ou o mix de todos. As especificações incluem: RDF, XML,TRiX, ATOM, YADS (*OPEN ARCHIVES INITIATIVE*, 2007).

Tim Berners-Lee em 2006 concebeu o*Linked Data*como sendo uma tecnologia via Web capaz de conectar dados relacionados, conferindo maior empregabilidade e eliminando obstáculos a uma maior integração destes mesmos dados (*LINKED DATA*,2013). O*Linked Open Data* (LOD) iniciou-se como um modelo aberto e comunitário em2007 tendo como foco a publicação de blocos de dados intrinsecamente interligados onde a*World Wide Web Consortium* (W3C) figurava como fomentadora na fase de projeto.

Com o avanço nos processos de interoperabilidade e dedados abertos vinculados passa a ser viável um formato onde a informação científica é difundida na Web e captada por indistintos usuários conferindo acréscimo deimportância aos sistemas de informação (HEATH; BIZER, 2011). Assim como no caso do LOD, a tecnologia OAI-ORE evoluiu de uma Web marcada essencialmente por textos e documentos, ausentes de semântica culminando em 2001, com as primeiraspropostas desta por Tim Berners-Lee (BERNERS-LEE; HENDLER; LASSILA, 2001).

Diferentemente de outrora, verifica-se na atualidade quenão são poucas as iniciativaspara um efetivo crescimento de *datasets* do LODno formato *Resource Description Framework* (RDF) – o modelo de dados da Web semântica(*LINKED DATA*, 2013). Assim, o objetivo deste estudo é o de abordar as tecnologias OAI-ORE e LOD com fins de verificar suas características e complementaridades buscando responder o seguinte questionamento: Quais são as peculiaridades existentes entre as tecnologias OAI-ORE e LOD?

O método de análise de conteúdo empregado no presente trabalho foi o delineado por Bardin (1979) que, conceitua a técnica como sendo um conjunto de procedimentos de análise dos documentos visando obter, por artifícios sistemáticos e objetivos de descrição dos textos informações que, permitam a inferência de conhecimentos relativos às condições de produção

ISBN 978-85-61115-09-8

e recepção (variáveis inferidas). Contextualizando o ato de pesquisar temos que, para Silva e Menezes (2001), se refere a um conjunto de procedimentos racionais e sistemáticos que tem por desígnio buscar soluções aos problemas que são propostos.

A pesquisa científica, de forma ampla, trata de estudos projetados e desenvolvidos em conformidade às normas da metodologia científica, sobre um objeto ou uma conjuntura, onde para Ander-Egg *apud* Marconi e Lakatos (2006), esta se constitui num procedimento reflexivo sistemático, controlado e crítico, que consente desvendar fatos, dados, relações ou leis, em várias áreas do conhecimento. Desenvolve-se a pesquisa mediante o ingresso de conhecimentos específicos com o emprego de métodos, técnicas e demais procedimentos inerentes à atividade empírica. Conforme pontua Gil (2002) de fato, a pesquisa aprimora-se no decorrer de um processo que abarca várias etapas, partindo da formulação do problema até a exposição dos resultados.

A análise de conteúdo se deu em três momentos distintos: a pré-análise, a exploração do material e a interpretação das informações coletadas com sua subsequente transcrição. O corpus da investigação se constituiu de coletas de dados exclusivamente distribuídos ao longo das fontes primárias e secundárias de informação disponíveis, ou seja, periódicos científicos, documentos eletrônicos de entidades oficiais, bibliografia especializada, fóruns de debates e pesquisa acadêmica.

2.A Tecnologia OAI-ORE

O mais recente projeto da Iniciativa dos Arquivos Abertos (OAI),concebido em 2006, o OAI-ORE, surge da necessidade de se integrar repositóriosde imagens, textos, áudios, gráficos e vídeos em seus vários formatos, para quepossam ser disponibilizados como resultado à pesquisa.

O projeto, conforme a Open Archives Initiative (2008) coloca, foi fomentadopor grandes instituições: *Andrew W. Mellon Foudation, Microsoft Corporation, Coalition for Networked Information* e *Digital Library Federation*.As considerações finais tratam do fechamento do tema, ainda que reconhecendo os limites do próprio artigo para apontar soluções, podendo-se pontuar a necessidade de novas investigações.Abordando a questão da interoperabilidade, Lagoze e Van de Sompel (2001),trazem a luz, a percepção de que esta, apresenta inúmeras facetas, compostas deconjuntos de metadados, protocolos de acesso, formatos de documentos enomenclaturas uniformes.

No caso da tecnologia OAI-ORE, seu funcionamentoocorre por meio de especificações que permitem aos repositórios, trocar informaçõessobre os objetos digitais que os compõem.A representação desses objetos digitais promove o acesso, extração e oreagrupamento das informações, mudando a concepção formada a respeito dosrepositórios enquanto meros depósitos de informações estáticas. As "normas" OAI-OREprovêem as fundações para aplicações e serviços com o intuito de visualizar,preservar, transferir, sumarizar e melhorar o acesso ao conteúdo disperso na Web,incluindo documentos com várias páginas, com múltiplos formatos em repositóriosinstitucionais e privados de acesso aberto e anexos de dados científicos(*DATAGAZETTEER*, 2008).

O termo objeto digital complexo, conforme pontuam Bekaert *et al.*(2006), diz respeito a uma generalização de ações em várias áreas,que procuram estabelecer modelos de interoperabilidade para compartilhamento deconteúdo digital. Estes objetos são tidos como unidades essenciais de reuso nocontexto de bibliotecas digitais, sendo habitualmente compostos por conteúdo emetadados que os delineiam e organizam internamente.

Conforme Lagoze *et al.* (2008) definem, o OAI-ORE reformula a noção derepositório orientado a objeto digital com agregação limitada de recursos web.Fazendo com que o

ISBN 978-85-61115-09-8

conteúdo da biblioteca digital seja mais integrado com aarquitetura da web e, portanto, mais acessível aos aplicativos informáticos.Voltada às modalidades *eScience* e *eScholarship*, onde o conteúdo édistribuído através de múltiplos serviços e bases de dados, a tecnologia OAI-OREamplia os esforços de interoperabilidade descrevendo especificações de mineraçãode dados e reutilização de objetos digitais (LAGOZE *et al.,* 2008). A constituição deobjetos complexos, normalmente abrange a etapa de produção de conteúdo e a consequente acepção dos metadados. O acondicionamento desses elementos embibliotecas digitais possibilita sua padronização, por meio deoperações de recuperação da informação.A sistemática da interoperabilidade surge a partir do momento em que estesobjetos digitais rompem os contornos da biblioteca, com o propósito de se prestaremao reuso com finalidades diversas e em outras unidades. A partir daí se reconhece acomplexidade destes objetos digitais que são formados de diferentes formatos deimagens, sons e textos com extensões díspares, a questão da interoperabilidadepassa a ser muito mais significativa que uma simplória alternância de documentosentre repositórios (MARCONDES; SAYÃO, 2001).Não se trata apenas de resolver uma demanda informacional simples com arecuperação de um único objeto composto por um single format (DOC, RDF, PDF,MPEG, JPG, AVI, MP3).

A questão envolve a reutilização de conteúdo formado demídias conjugadas, o que precisamente suplanta as expectativas do usuário noprocesso de recuperação da informação. No caso de um exemplo prático deoperação de busca e recuperação de informação, pode-se citar o caso hipotético deum pesquisador que deseja material a respeito de um determinado artista com váriosmateriais distribuídos em diferentes suportes (PDF, JPEG, MP3), onde constamobras literárias, cinematográficas e de musicalidade. Com a intervenção datecnologia OAI-ORE, tudo a respeito desta personalidade será coletado entre osrepositórios, recuperado em seus vários formatos (vídeo, texto e áudio) e mantidoem um repositório criado para esta demanda específica, estando, portanto, prontopara a sua reutilização. Ações de padronização destinam energia com estepropósito, ao deliberar arquétipos que têm como fundamento, formatos abertos demetadados e provêem formas de interoperabilidade e reuso.O pressuposto básico para que a tecnologia de troca e reuso dos objetosdigitais (ORE) possa atuar é o de os repositórios informacionais participarem dainiciativa dos arquivos abertos, disponibilizando seus conteúdos livremente. Quanto às agregações, segundo Cavalcante (2007), são junções em que um objeto é parte de outro, de maneira que a fração pode viver sem o todo. Consiste de um objeto contendo menções para outros objetos, de forma que o primeiro se caracterize por "todo", e que os objetos referenciados sejam as partes. Habitualmente emprega-se agregação para ressaltar minúcias de uma futura implementação.

Concebe a perspectiva de que um objeto usa outro sem ser titular dele e, assim não é responsável pela sua concepção ou extermínio. Os recursos podem ser compreendidos como os Uniform Resource Locators ou endereços eletrônicos na Web e seu mapeamento consiste na varredura de determinadas relações estabelecidas entre alguns destes Localizadores-Padrão de Recursos. No caso da tecnologia *Open Archives Initiative – Object Reuse and Exchange*, os mapas de recursos são as descrições destas agregações de objetos digitais já incorporados.

3.A Tecnologia LOD e algumas similaridades com o OAI-ORE

O *Linked Open Data* (LOD) foi criado essencialmente como uma ação para publicação de dados na Web intencionando cooperar para a organização e recuperação da informação na rede mundial. Assim como a internet tem proporcionado mudanças de paradigmasna forma como acessamos informações e mais especificamente os documentos que contém estas, também ela pode revolucionar a maneira como descobrimos e interagimos no meio virtual, integrando e concebendo vinculações entre os dados digitados a partir de fontes diferentes.

Estes mesmos dados em sua grande maioria encontram-se extremamente divergentes em bases de dados pulverizadas ao longo de uma rede infindável de conexões oriundas de

ISBN 978-85-61115-09-8

diferentes localidades, ou simplesmente residem em sistemas heterogéneos dentro de uma única organização que, em muitos casos desconhece na integralidade sua constituição.

Assim, como na tecnologia OAI-ORE, grandes instituições financiaram o projeto LOD e conforme aponta Baptista(2002), entre estes ainda figura o W3C criado em 1994, com o objetivo dedesenvolver protocolos comuns para promover a evolução e assegurar a interoperabilidade da Web. Reside ai outra similaridade com o *Open Archives Iniciative* que criou o protocolo PMH (*Protocol for Metadata Harvesting*) com fins de promoção da interoperabilidade.

Ainda quanto à questão dos financiadores, o W3C tem como mantenedores o: *MIT (Massachussets Institute of Technology); KEIO University, ERCIM (European Research Consortium for Informatics)* e o *Beihang University (Beijing University of Aeronautics & Astronautics).* Tecnicamente, o termo LODdiz respeito a dados publicados na internet de tal forma que são legíveis por máquina, o seu significado é explicitamente definido como ligado a outros conjuntos de dados externos de forma extensiva. Enquanto as unidades primárias da Web de hipertexto são constituídas pelos HTMLs (*Hyper Text Markup Languages*), ou seja, documentos conectados por *hiperlinks*, o LOD conta com dados em formato RDF (*Resource Description Framework*), conforme definem Klyne e Carroll, (2004).

Berners-Lee*et al.* (2006) concebeu um agrupamento de normas para a publicação de dados na Web de maneira que todos estes se tornam-se parte de um espaço único global, instituindo:Uso de URIs como resultados para os itens; Uso de HTTPs para consultasdesses resultados; Uso do padrão RDF no fornecimento de informações; Inclusão de sentenças RDF com *links* para outras URIs. O RDF é apropriado para propagar descrições a respeito derecursos, incrementando o processamento automatizado por máquinas.

Conforme Manola e Miller (2004), o processamento automatizado demanda um sistema automatizado que identifique o sujeito, o predicado e o objeto em uma declaração sem ambiguidade, além de uma linguagem para representar estas descrições que facilite o intercâmbio de informaçõesentre máquinas. De acordo com as descriçoes de Lassila e Swick (1998), além de Manola (2004), a publicação dos dados é realizada no formato RDF, com representação em grafos genéricos na forma de triplas – sujeito, predicado, objeto – nas quais o sujeito e o objeto são recursos identificados por URIs. Por sua vez, o predicado especifica como o sujeito e o objeto estão relacionados, sendo também representado por uma URI (BRICKLEY; GUHA, 2004).Alimentando a expressão RDF, permanece um conjunto de triplas, cada uma integrada de um sujeito, um predicado e um objeto, onde as triplas, que representamafirmações sobre relacionamentos entre entidades concebidas por séries de nós,são denominadaspor grafos de modelos de dados. Segundo Gauthier (2006), cada nó pode ser uma URI, um literal ou estar em branco (não podendo ser identificado) eos predicados são identificados por URIs e podem ser também nós em um grafo.

4. Características e Complementaridades

O LOD se constitui como um dos projetos mais notáveis no campo da Web Semântica sendo apoiado pela W3C *Semantic Web Education and Outreach Interest Group*(2006), cujo principal objetivo é identificar e extrair dados de bases abertas existentes (como banco de dados relacionais abertos, arquivos XML, páginas HTML, etc.), convertê-los para o formato recomendado, interligá-los e publicá-los na Web para uso público (BIZER *et al.,* 2009).

Do ponto de vista tecnológico, o *Linked Open Data* ou simplesmente LOD, tem como questões principais a interoperabilidade e a integração de dados em larga escala.Desenvolvida com fins acadêmicos de pesquisa e aprendizagem, atecnologia OAI-ORE é uma norma que se diferencia do LOD em função desta primeira se prestar declaradamente à pesquisa, o que não exclui a referida aplicabilidade quanto ao *Linked Open Data*. Afinal, opróprio termo o define como sendo um estilo comunitário de se publicar einterligar dados estruturados na Web, onde

ISBN 978-85-61115-09-8

o objetivo central é o de permitir que as pessoas compartilhem dados estruturados de maneira simplificada.

Quanto mais um dado for interligado com outros, maior é o seu valor e sua utilidade e entre os benefícios proporcionados por esta tecnologia, constam:

• Os usuários podem encontrar URIs em formato RDF obtendo informação adicional;

• Informações de diferentes origens com a criação de links em RDF;

• O modelo permite a unificação de padrões;

• Combinalinguagens de esquemas como RDFS ou OWL, em dados semiestruturados;

• Como aWeb de dados é aberta, novas fontes surgem em tempo real via*links*.

Sendo assim, há complementaridade nas ações entre o OAI-ORE e o LOD, implicando em que, enquanto um recupera uma gama de formatos diferenciados de acordo com a demanda apresentada pelo usuário o outro organiza e publica de forma aberta e colaborativa tais recursos.

Aplicações genéricas para oLOD permitem o consumo de dados relacionadosa múltiplos domínios distribuídos pelo amplo espaço de dados global, desta forma, aopercorrer os *links* em RDF é possível explorar e descobrir novas informações na Web.

5. Considerações Finais

Frente à insuficiência financeira por qual passam as agências de fomento àpesquisa dos países em desenvolvimento, a hipótese de redução de custos é algo aser considerado, pois, isto admite que informações relevantes tornem-se disponíveispor metodologias e tecnologias *open source*. Segundo Hexsel (2002), a prerrogativamais significativa do código aberto é possibilitar ao usuário fugir da tecnologiaproprietária.

O presente trabalho se baseouno estudo de literatura orientada a objetosdigitais reutilizáveis e intercambiáveis, produzida por especialistas pesquisadores einstitucionais. São vislumbradas aplicações nas mais diversas atividades depesquisa que se utilizam dos repositórios de vídeos, fotos, imagens em geral,bancos de dados da química, física, bibliotecas virtuais ou digitais de matemática,etc.

Com o advento do ciberespaço, as unidades de informação são de extremanecessidade para gerenciar e prover o crescente número de itens digitais dispostosaleatoriamente. A possibilidade de reutilização destes objetos digitais compostosabre uma imensa via de recursos e possíveis aplicações (principalmente emrepositórios institucionais), constituindo-se, portanto, em uma estrutura emergentede agregações flexíveis na gestão informacional.

Empregando estruturas de acesso padronizados disponibilizados abertamente como o LOD, o acesso às fontes de dados pode ser ilimitado, aproveitando toda a potencialidade daweb. A quantidade de dados recuperados com atecnologia LODprospecta um crescimento muito rápido cobrindo os mais variados domínios se conjugada ao OAI-ORE.

As aplicações para estes dados são de umpotencial muito grande de revolucionar a maneira de como tais dados podem ser consumidos, com fins acadêmico-científicos ou empresariais, revertendo ao país frutos econômicos relevantes.Porém ainda há muito a ser feito no âmbito da pesquisa e este estudo espera fomentar tais iniciativas.

Observa-se que pode haver uma interação entre as duas tecnologias, pois enquanto uma recupera tudo o que há sobre um determinado assunto na Web em seus variados formatos, a outra se ocupa em reestruturá-los de forma que tais resultados sejam organizados de acordo com a sua relevância.

ISBN 978-85-61115-09-8

Portanto, as duas sistemáticas são importantes para a evolução da recuperação da informação, sobretudo a de valor científico e empresarial.

Em suma, o LOD se aplica com eficiência aos dados que são ligados por essência, tais como as redes de toda ordem, entre essas as sociais (tão em voga) e as informações sobre determinados assuntos (DBPedia), sendo assim, em conjunto com as possíveis aplicações do projeto OAI – ORE, é possível a implementação de um volume maior de novos recursos aos repositórios institucionais e não apenas etão somente, uma maior interoperabilidade.

O potencial dos repositórios digitais online direcionados para oensino/aprendizagem é vasto, especialmente aqueles em que os conteúdos sãodisponibilizados livremente e sem custos de utilização, como meio de promoção daeducação e aprendizagem ao longo da vida (OLCOS, 2007).

Tais aplicações são consoantes com os ideais propostos por OCDE (2008)pontuando que, a definição mais clara para o que são os recursos educacionaisabertos, versa se tratarem de materiais digitalizados oferecidos livre e abertamente aprofessores, alunos e autodidatas para utilização e reutilização no ensino,aprendizagem e investigação.

Constitui-se, portanto, mais uma importante estruturade agregações flexíveis de gestão de objetos digitais, principalmente os derepositórios institucionais.

Até o final do ano de dois mil e oito, os pesquisadores, Carl Lagoze eHerbert Van de Sompel, lideraram pessoalmente, o desenvolvimento do projeto OAI– ORE.

Após a fase de projeto, realizaram a demonstração do produto final aopúblico e muito embora os idealizadores estivessem sempre acompanhando oprogresso de iniciativas autônomas por meio de fóruns de debates e listas dediscussões, não mais se envolveram diretamente na arquitetura do *software*.Designando que, as contribuições surgissem de quem tivesse anecessidade.

O financiamento cessou e a norma ORE, desde então, criou vidaprópria nas mãos de uma legião de desenvolvedores.

O ambiente colaborativo da iniciativa de arquivos abertos possibilita o acessoao código fonte e a partir daí novas agregações podem surgir espontaneamente,em sintonia com as necessidades das instituições, unidades informacionais ecomunidade científica.

Referências

Baptista, A. A. R. P. Informática online: um enquadramento para a publicação em linha de revistas científicas electrónicas (**Tese de Doutoramento**). Universidade do Minho, Guimarães, 2002.

BARDIN, Laurence. **Análise de conteúdo.**Lisboa: Edições 70, 1979.

BEKAERT, J.; DE KOONING, E.; VAN DE SOMPEL, H. Representing digitalassets using MPEG-21 digital item declaration. **International Journal on Digital Libraries**, 6(2), pp. 159-173. 2006. Disponível em: <http://public.lanl.gov/herbertv/papers/Papers/2006/IJDLbekaert.pdf>. Acesso em: 29set. 2014.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. TheSemantic Web. **Scientific American**, p. 34-43, May2001.
Disponível em: <http://www.scientificamerican.com/article.cfm?id=the-semantic-web>. Acesso em: 29 set. 2014.

ISBN 978-85-61115-09-8

BERNERS-LEE, T.; CHEN, Y.; CHILTON, L.;CONNOLLY, D.; DHANARAJ, R.; HOLLENBACH,J.; LERER, A.; SHEETS, D. Tabulator: exploringand analyzing linked data on the semantic web.In: **INTERNATIONAL SEMANTIC WEB USER INTERACTION**, 3. 2006. Disponívelem: <http://swui.semanticweb.org/swui06/papers/Berners-Lee/Berners-Lee.pdf>. Acesso em: 29set. 2014.

BRICKLEY, D.; GUHA, R. V. 2004. **RDF vocabulary description language 1.0:** RDF Schema. W3C Recommendation.Disponível em: <http://www.w3.org/TR/rdf-schema/.>. Acesso em: 30 out. 2014.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. LinkedData: the story so far. **International Journal on Semantic Web and Information Systems**, v. 5, n. 3,p. 1-22, 2009. Disponível em: <http://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linkeddata.pdf>. Acesso em: 29set. 2014.

DATAGAZETTEER. Open archives initiative announces public beta release ofobject reuse and exchange specifications. **Blog LISWire The Librarian's News Wire.** 2008. Disponível em: <http://liswire.com/node/81>. Acesso em:29 de set.2014.

GAUTHIER, F. A. O. RDF e RDFShema. UFSC/EGC, 2006. (Material de Aula). Disponível em: <http://www.inf.ufsc.br/~gauthier/EGC6006/material/Aula%202/RDFRDFSchema.pdf>. Acesso em: 31 out. 2014.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa.**4 ed. São Paulo: Atlas, 2002.

HEATH, T.; BIZER, C. **Linked Data:** envolving the web into a global data space. California, USA: Morgan &Claypool, 2011.

HEXSEL, Roberto A. **Propostas de ações de governo para incentivar o uso de software livre.** Relatório Técnico do Departamento de Informática da UFPR,004/2002, out 2002. Disponível em:<http://www.inf.ufpr.br/info/techrep/RT_DINF004_2002.pdf>. Acesso em: 29 set.2014.

KLYNE, Grahan; CARROLL, Jeremy J.**Resource description framework:** concepts and abstract syntax, 2004. Disponível em: <http://www.w3.org/TR/rdf-concepts/>. Acesso em: 29 set. 2014.

LAGOZE, C. et al. **Object Re-Use & Exchange:** a resource-centric approach. Arxiv preprint. 2008. Disponível em: <http://arxiv.org/ftp/arxiv/papers/0804/0804.2273.pdf>. Acesso em: 29set. 2014.

LAGOZE, C.; VAN DE SOMPEL, H. The open archives initiative: building a lowbarrier interoperability framework. In: **JCDL - Joint Conference on Digital Libraries**,01, 2001, Roanoke, Va. Proceedings of the First ACM/IEEE-CS Joint Conference onDigital Libraries. New York: ACM Press, 2001. p. 54 - 62. Disponível em: <http://www.openarchives.org/documents/jcdl2001-oai.pdf>. Acesso em: 29set. 2014.

LASSILA, O.; SWICK, R. R. **Resource Description Framework (RDF):**model and syntax specification, 1998.

ISBN 978-85-61115-09-8

LINKED DATA. Disponível em: <http://www.linkeddata.org>. Acesso em: 29 set. 2014.

MANOLA, F.; MILLER, E. 2004. **RDF primer.**W3C Recommendation. Disponível em: <http://www.w3.org/TR/rdf-primer/.>. Acesso em: 30 out. 2014.

MARCONDES, C. H.; SAYÃO, L. F. Integração e interoperabilidade no acesso arecursos informacionais eletrônicos em C&T: a proposta da biblioteca digitalbrasileira. **Ci. Inf.**, Brasília, v.30, n.3, p.24-33, set./dez. 2001. Disponível em: <http://revista.ibict.br/ciinf/index.php/ciinf/article/view/190/167>. Acesso em: 29 set.2014.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Técnicas de pesquisa:** planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração, análise e interpretação de dados. 6 ed. São Paulo: Atlas, 2006.

OCDE. **El Conocimiento libre y los recursos educativos abiertos.** Paris: OCDE,tradução de Junta de Extremadura, 2008. Disponível em: <http://www.oecd.org/dataoecd/44/10/42281358.pdf>. Acesso em: 29 set. 2014.

OLCOS. **Open educational practices and resources:** OLCOS roadmap 2012. EUfunded OLCOS. 2007. Disponível em: <http://www.olcos.org/cms/upload/docs/olcos_roadmap.pdf>. Acesso em: 29set.2014.

OPEN ARCHIVES INITIATIVE. **Open archives initiative object reuse and exchange:** ore specifications and user guides primer. 2008. Disponível em: <http://www.openarchives.org/ore/1.0/primer>. Acesso em: 29 set. 2014.

OPEN ARCHIVES INITIATIVE. **Open archives initiative object reuse and exchange:** the OAI-ORE effort: progress, challenges, synergies. JCDL: Vancouver, 2007. Disponível em: <http://www.openarchives.org/ore/documents/orejcdl2007.pdf>. Acesso em: 29 set. 2014.

SILVA, E. L.; MENEZES, E. M. **Metodologia da Pesquisa e elaboração de dissertação.** Florianópolis: Laboratório de Ensino a Distância da UFSC, 2001.

VAN DE SOMPEL, H.; LAGOZE, C. The Santa Fé convention of the open archives initiative. **D-Lib Magazine**, v. 6, n. 2, 2000. Disponível em: <http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html>. Acesso em: 29set. 2014.

WORLD WIDE WEB CONSORTIUM. 2006. Disponível em:<http://www.w3.org>. Acesso em: 29set. 2014.

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

AVALIAÇÃO DA PRODUÇÃO CIENTÍFICA SOBRE *ENTERPRISE LINKED DATA*

Rafael de Moura Speroni
rafael@ifc-araquari.edu.br

Evelin Priscila Trindade
evelin.trindade@gmail.com

Marco A. Neiva Koslosky
marco@ifsc.edu.br

Marcelo Macedo
marcelomacedo@egc.ufsc.br

Fernando A. Ostuni Gauthier
gauthier@egc.ufsc.br

Resumo
Linked Data ou Dados Ligadossão uma forma de publicar documentos e dadosinter-relacionados na web. As técnicas envolvidas tornaram-se interessantes para as organizações por moldarem uma solução geral para a divulgação e integração de informações (WOOD, 2010). O termo *Enterprise Linked Data* refere-se à adoção dos princípios de *Linked Data*nas empresas, onde os sistemas de informação corporativos podem ser vistos como um espaço de *Linked Data (SERVANT, 2008).* Considerando que a adoção destes princípios é recente nos ambientes empresariais, este estudo tem por objetivo avaliar a produção científica relativa ao tema. Foram analisados 46 artigos, selecionados de um total de 165 publicações entre os anos de 2010 e 2014, quanto à (i) definição do termo *Enterprise Linked Data*, (ii) volume de publicações por ano, (iii) principais fontes de publicação, (iv) autores, instituições, países e(v) principais referências bibliográficas. Conclui-se com essa pesquisa que ainda existe muito sobre o termo *Enterprise Linked Data* para se explorar, tanto na academia como pelas empresas e os dados apresentados sobre o tema são de grande relevância para pesquisas futuras.

Palavras-chave: Enterprise Linked Data. Revisão Bibliométrica. Dados Ligados. Dados Empresariais.

Abstract
Linked DataorLinkedData isa mannerof publishingdocumentsand interrelatedwebdata.These techniques had becomeinterestingfor organizationsbyshapinga general solutionfor the disseminationand integration ofinformation(Wood, 2010). The termEnterpriseLinked Datarefers to theadoption of the principlesofLinked Datain companieswherecorporateinformationsystemscan be seenas aLinked Data space (SERVANT, 2008). Whereasthe adoptionof these principlesisnewinenterprise environments, this study aims toevaluate thescientific literatureon the subject. The analysis was performed from 46 articles, selected from a total of 165 publications between the years 2010 and 2014, regarding (i) definition of the term Enterprise Linked Data, (ii) volume of publications by year, (iii) major sources of publication, (iv) authors, institutions, countries, and (v) primary references. We conclude from this survey that there is still much about the term Enterprise Linked Data to explore, both in academia as by companies and data presented on the topic is of great relevance for future research.

ISBN 978-85-61115-09-8

Keywords: Enterprise Linked Data. Bibliometric review. Linked Data. Enterprise Data.

1. Introdução

A adoção das tecnologias de *web* semântica tem levado à criação de um espaço global de dados. As iniciativas da comunidade *Linking Open Data*[37] contribuíram muito para a concretização da *web* de dados, descrevendo melhores práticas (BERNERS-LEE, 2006; BIZER, CYGANIAK; HEATH, 2007; HEATH; BIZER, 2011), publicando grandes conjuntos de dados em RDF na *web*(BIZER *et al.*, 2009; VILCHES-BLÁZQUEZ *et al.*, 2010) e, consequentemente, dando início a uma era de novas possibilidades de uso para estes dados (SERVANT, 2008).

Embora muitas aplicações de *Linked Data* tenham sido desenvolvidas sobre dados de acesso público, o reconhecimento dos benefícios destas tecnologias passou a despertar o interesse das empresas, cujos sistemas de informação também podem ser vistos como espaços de dados ligados. Para Hu e Svensson (2010), ainda que os dados empresariais apresentem diferenças significativas dos dados públicos na internet, uma vantagem evidente das tecnologias semânticas é a integração de conjuntos de dados distribuídos, beneficiando as companhias com um grande valor de retorno.

Pesquisas recentes sobre *Enterprise Linked Data* trazem informações importantes sobre este padrão para publicação de dados organizacionais na *web*, apresentando resultados melhores na integração e recuperação dos dados publicados. Apesar de seu potencial, a adoção das técnicas de *web* semântica acontece de forma mais lenta nas empresas. As explicações passam pela preocupação das empresas com a segurança e controle de acesso aos dados (ORTIZ *et al.*, 2013), ou pela possibilidade de que tais tecnologias sejam vistas como promessas, ou modismos tecnológicos, e não esteja suficientemente claro o que acrescentam ao cenário existente(SERVANT, 2008).

A aplicação dos princípios de *Linked Data* em dados empresariais é referenciada na literatura pelos termos "*Enterprise Linked Data*", "*Linked Enterprise Data*" ou "*Linking DataEnterprise*". Valendo-se de uma revisão bibliométrica, o objetivo desse trabalho é verificar o volume de publicações por ano, principais fontes de publicação, autores, instituições, países e as principais referências bibliográficas.

O artigo está dividido de acordo com as seções a seguir: na seção dois é apresentada a descrição dos procedimentos metodológicos utilizados para o desenvolvimento da revisão sistemática; na seção três são apresentados os resultados obtidos com a pesquisa e na quarta seção são apresentadas as considerações finais.

2. Procedimentos metodológicos para a realização da pesquisa bibliométrica

Este estudo descreve a avaliação da produção científica sobre "Enterprise Linked Data". Segundo Pizzani et al. (2008, p. 69-70), "um dos mecanismos mais utilizados pela comunidade científica para a disseminação dos resultados das pesquisas é a publicação de artigos científicos em revistas, os chamados periódicos científicos e, para avaliar a produção científica de um determinado grupo de pesquisa foram elaborados indicadores para medir a sua visibilidade científica".

Nesse sentido, como forma sistêmica para obtenção e análise dos dados foi realizada uma bibliometria focada nos artigos científicos referenciados em bases internacionais de

[37] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

ISBN 978-85-61115-09-8

pesquisa.Conforme Araújo (2006, p.12) a bibliometria consiste "na aplicação de técnicas estatísticas e matemáticas para descrever aspectos da literatura e outros meios de comunicação (análise quantitativa da informação)". A bibliometria, como método quantitativo de investigação da ciência, utiliza a análise de citações como uma de suas ferramentas como forma de medir o impacto e a visibilidade de determinados autores dentro de uma comunidade científica, verificando quais escolas do pensamento vigoram dentro das mesmas (VANZ; CAREGNATO, 2003).As principais leis bibliométricas utilizadas neste trabalho foram: Lei de Bradford (produtividade de periódicos), Lei de Lotka (produtividade científica de autores) e Lei de Zipf (frequência de palavras).

2.1 Etapa 1 – Consulta nas bases de dados científicas

A busca foi realizada em três bases de dados científicas: *Scopus*, *Web of Science* e *IEEE Xplore Digital Library*. As bases escolhidas tem caráter multidisciplinar e permitem a exportação dos resultados da busca para ferramentas de gerenciamento bibliográfico.

A *Scopus* é uma base científica que indexa mais de 21.000 títulos e 53 milhões de registros com atualização diária. A *Web of Science* conta com mais de 90 milhões de registros, sendo considerada uma relevante base para estudos bibliométricos(BRAMBILLA; STUMPF, 2012). O *IEEE Xplore* indexa conteúdos técnicos e científicos pubicados pelo IEEE (*Institute of Electrical and Electronics Engineers*) e seus parceiros. Seu conteúdo explora *journals*, conferências, padrões técnicos, *e-Books* e cursos educacionais, com aproximadamente 25.000 novos documentos adicionados por mês.

Os termos utilizados para a busca foram "*Enterprise Linked Data*", "*Link* Enterprise Data*" e '(("*Linked Data*") and *Enterprise*)'. A utilização do caractere "*" garante que outras palavras sejam encontradas, como "*Linked*" e "*Linking*", aumentando a abrangência da busca. A busca foi realizada em setembro de 2014 e a delimitação foi de cinco anos, buscando-se obras publicadas de 2010 a 2014.

2.2 Etapa 2 – Exportação e seleção dos dados

A segunda etapa tem por objetivo selecionar o corpo de documentos a ser analisado. Para isto, os resultados das buscas foram exportados e carregados, na ferramenta de gerenciamento de referências EndNote, em um conjunto único de documentos. Com base nos dados carregados, o software possui recursos que auxiliam a sua padronização. O primeiro critério de seleção aplicado foi a exclusão de obras sem autor. O segundo critério aplicado foi a exclusão de artigos duplicados, em virtude da eventual ocorrência dos mesmos em mais de uma base buscada.

O terceiro critério de exclusão foi a disponibilidade de texto completo. Foram excluídos os artigos cujos textos completos não estavam disponíveis gratuitamente para download, cuja extensão do documento não era superior a uma página ou que estavam fora do contexto deste estudo, por não tratarem da aplicação de *Linked Data* a dados empresariais. Para a identificação de documentos fora do escopo procedeu-se a leitura dos resumos dos artigos.

ISBN 978-85-61115-09-8

2.3 Etapa 3 – Padronização dos dados

A terceira etapa consiste na padronização dos dados. Embora as bases científicas permitam a exportação para o mesmo formato, há pequenas diferenças nos dados, como os nomes de autores abreviados ou por extenso.

Além da padronização, foram acrescentadas aos dados as coordenadas geográficas das instituições de vínculo dos autores dos artigos. Estas coordenadas foram obtidas manualmente com o auxílio da ferramenta *Google Maps,* e possibilitam a geolocalização dos autores.

2.4 Etapa 4 – Análise dos dados e redação do documento

Com os dados padronizados foi possível passar à etapa de análise. Buscou-se identificar: a definição do termo *Enterprise Linked Data*, quantidade de publicações por ano segundo o tipo de publicação, as principais fontes, os principais autores, suas instituições e os países, e as principais referências bibliográficas utilizadas.

3.Análise dos documentos encontrados

Esta seção apresenta os resultados das análises feitas a partir do corpo de documentos selecionado segundo os critérios estabelecidos nos procedimentos metodológicos.

3.1 Dados gerais sobre a busca

A busca realizada nas três bases escolhidas retornou um total de 165 registros que, após a aplicação dos critérios de seleção, resultaram em 46 artigos selecionados para análise. A Tabela 2 apresenta os números relativos à seleção dos artigos.

Tabela 2 ***- Quantidade de documentos selecionados***

Base de Dados	Resultado da busca	Sem Autor	Duplicados	Sem texto completo ou fora de escopo	Total
IEEE Xplore	31	01	17	04	**09**
Scopus	115	12	05	61	**37**
Web of Science	19	00	16	03	**00**
Total	**165**	**13**	**38**	**68**	**46**

Fonte: elaborado pelos autores

Dos 46 artigos selecionados, cinco são capítulos de livros, 36 são artigos em conferências e cinco são artigos em *journals*. Esses trabalhos foram escritos por 146 autores de 55 instituições em 15 países diferentes. Os artigos foram indexados nas bases de dados científicas,utilizando 527 palavras-chave e utilizam um total de 903 referências. A tabela 2 apresenta os documentos analisados.

ISBN 978-85-61115-09-8

Tabela 2 - ***Documentos analisados***

Ano	Artigos
2010	(ALLEMANG, 2010; FOGAROLLI; BOUQUET, 2010; HALB *et al.*, 2010; HARRIS, 2010; HARRIS, ILUBE; TUFFIELD, 2010; HU; SVENSSON, 2010; HYLAND, 2010; PASSANT *et al.*, 2010; ROHDE; SUNDARAM, 2010)
2011	(ÁLVAREZ *et al.*, 2011; CHRISTIDIS, MENTZAS; APOSTOLOU, 2011; FENG *et al.*, 2011; GAO, DERGUECH; ZAREMBA, 2011; GRAUBE *et al.*, 2011; HASSANZADEH *et al.*, 2011; KIRRANE, 2011; LANTHALER; GUTL, 2011; PAGANO, 2011; SACCO, PASSANT; DECKER, 2011; SLEIMAN, RIVERO; CORCHUELO, 2011; WESTERSKI; IGLESIAS, 2011)
2012	(ALVAREZ *et al.*, 2012; ASSAF *et al.*, 2012; CHENG *et al.*, 2012; CURRY, 2012; THOMA, SPERNER; BRAUN, 2012; TUÁN *et al.*, 2012; UMBRICH *et al.*, 2012; WURZER, 2012)
2013	(BHIRI, DERGUECH; ZAREMBA, 2013; BIANCHINI, DE ANTONELLIS; MELCHIORI, 2013; CURRY *et al.*, 2013; HOFFMANN *et al.*, 2013; HOVER; MUHLHAUSER, 2013; NASEER, LAERA; MATSUTSUKA, 2013; ORTIZ *et al.*, 2013; RUSITSCHKA *et al.*, 2013; SALMEN *et al.*, 2013; SERRANO *et al.*, 2013; SHAOPENG, JIANHUI; ZHIHONG, 2013; THOMA *et al.*, 2013)
2014	(ABBAS; OJO, 2014; ALOR-HERNÁNDEZ *et al.*, 2014; CORRY *et al.*, 2014; OMITOLA *et al.*, 2014; WEICHSELBRAUN, STREIFF; SCHARL, 2014)

Fonte: elaborado pelos autores

Na segunda etapa, os dados foram exportados e carregados no software Endnote, um gerenciador de bibliografias para publicação de documentos científicos. Essa ferramenta possibilita a padronização dos dados e conta com a verificação de artigos duplicados, bem como, a busca automatizada pelo texto completo. No software, ainda foram removidas as obras sem a indicação de autores.

Pela leitura dos resumos dos artigos, buscou-se a identificação daquelas obras que se encontravam fora do escopo da busca realizada, para que fossem removidas do corpo de documentos. Também foram desconsiderados aqueles trabalhos cuja extensão do documento não passava de uma página.

Os dados padronizados no software Endnote foram exportados para uma base de dados, onde foram adicionadas informações referentes à localização geográfica das instituições as quais os autores estavam vinculados, bem como o cadastro das referências bibliográficas de cada obra.

A quarta e última etapa foi a análise dos dados coletados a partir dos documentos encontrados. Esses dados podem ser vistos na próxima seção.

3.2 Definição do termo *Enterprise Linked Data*

Na análise dos artigos selecionados, identificou-se a forma como os autores conceituam o termo "*Enterprise Linked Data*". Embora seja perceptível o entendimento de que se trata da aplicação de *Linked Data* em empresas, poucos são os autores que preocupam-se com a definição.

Hu e Svensson (2010) caracterizam o termo "*linked enterprise data*" como um paradigma que surge em função do valor demonstrado pelas iniciativas de *linked data* que tiveram por objetivo melhorar a acessibilidade de dados a usuários principalmente públicos e acadêmicos.

Os autores destacam que:

> "As histórias de sucesso certamente não passaram despercebidas pelas grandes empresas. Tentativas cautelosas foram feitas para experimentar os princípios de *Linked Data* e avaliar seus benefícios, levando ao chamado paradigma '*linked*

ISBN 978-85-61115-09-8

> *enterprise data*', a contrapartida de *Linked Data* para o domínio dos negócios."(HU; SVENSSON, 2010)

No trabalho de Halb *et al.* (2010), embora não seja apresentada uma definição, são descritos três cenários possíveis para os quais as empresas podem adotar *linked data*:

- Interligar seus conteúdos, aumentando sua acessibilidade para humanos e máquinas;
- Integrar conteúdos de terceiros aos seus portais;
- Preparar seus próprios conteúdos para que terceiros possam utilizá-los, melhorando sua reusabilidade e visibilidade.

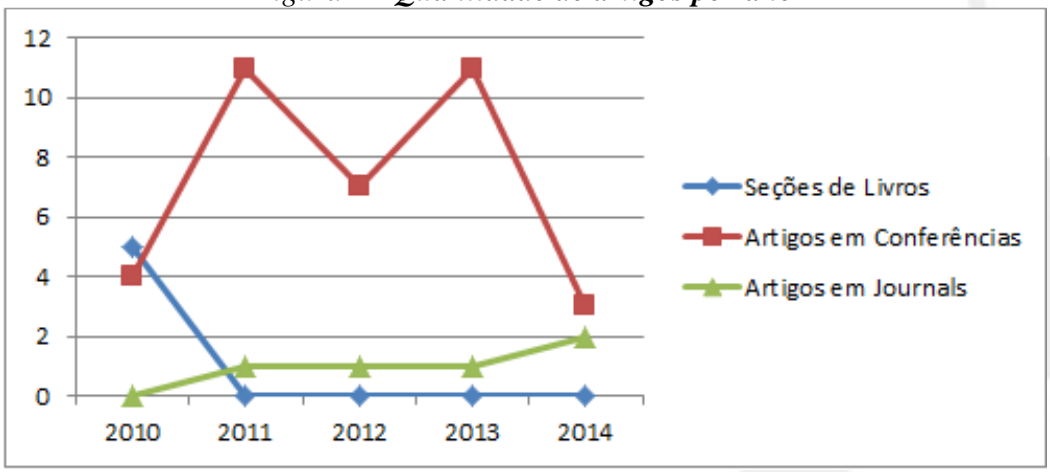
Sob o ponto de vista da empresa que utiliza *linked data*, Allemang (2010) a define como "*linked data enterprise*":

> "Uma organização na qual o ato da criação das informações está intimamente associado com o ato do compartilhamento da informação. Em uma analogia à *learning organization*, onde o aprendizado com uma atividade é tão importante quanto a atividade em si, em uma *linked data enterprise* o compartilhamento dos dados é tão importante quanto a sua criação." (ALLEMANG, 2010)

3.3 Publicações por ano

A distribuição do corpo de documentos analisados segundo o ano de publicação e tipo de obra é apresentada na Figura 1. Conforme é possível observar, as maiores incidências de publicação aconteceram nos anos de 2011 e 2013, contabilizando um total de 12 publicações em cada ano.

Figura 4 - ***Quantidade de artigos por ano***



Fonte: elaborado pelos autores

Nove obras foram publicadas no ano de 2010, sendo que cinco delas são seções de um mesmo livro: *Linking Enterprise Data*(WOOD, 2010), cujo editor é David Wood. Observa-se, ainda, que as obras selecionadas para análise foram, em sua maioria, artigos em conferências.

ISBN 978-85-61115-09-8

3.4 Principais fontes de publicação

Os artigos analisados foram publicados em 39 diferentes fontes de publicações. As fontes com maior número de publicações são apresentadas na

Tabela 3, e descritas em seguida.

Tabela 3- ***Principais fontes de publicação***

Fonte	Tipo	Quantidade de artigos
Linking Enterprise Data	Livro	5
Advanced Engineering Informatics	*Journal*	2
Hawaii International Conference on System Sciences (HCISS)	Conferência	2
International Conference o Semantic Systems (I-SEMANTICS)	Conferência	2
International Conference on Web Intelligence, Mining and Semantics (WIMS)	Conferência	2
International Semantic Web Conference (ISWC)		2

Fonte: elaborado pelos autores

A fonte mais representativa foi o livro "*Linking Enterprise Data*", publicado em 2010, que conta com cinco das publicações analisadas. O livro, editado por David Wood, apresenta algumas das primeiras aplicações em produção de *linking enterprise data*, e tem por objetivo servir como um roteiro para a replicação dos seus casos de sucesso. SegundoWood (2010), a parte I do livro provê orientações valiosas para aqueles que estão escrevendo casos de negócio, que precisam justificar os esforços de desenvolvimento interno, ou para quem necessita escrever propostas para fornecedores externos. A parte II provê material de assistência para gestores de negócio que desejam propor projetos de *Linked Data*.

Com dois artigos publicados, o "*Advanced Engineering Informatics*" é um *journal* publicado pela *Elsevier* que solicita artigos de pesquisa com ênfase em "conhecimento" e "aplicações de engenharia". O *journal* tem fator de impacto 2.068, e conta com quatro edições anuais.

A *"Hawaii International Conference on System Sciences*", com dois artigos publicados, é uma conferência internacional organizada pela *University of Hawai'i* e conta com o apoio da *IEEE Computer Society*. O evento está em sua 48ª edição e seus artigos são indexados na base da *IEEE Xplore*.

Com duas ocorrências, a "*International Conference on Semantic Systems – I-SEMANTICS*" é uma das maiores conferências na Europa nos campos de sistemas semânticos e *Web* Semântica. O evento conta com mais de 400 participantes todos os anos e encoraja pesquisas científicas e aplicações nas áreas Tecnologias Semânticas e *Linked Data*.

A "*International Conference on Web Intelligence, Mining and Semantics - WIMS*", onde foram publicados dois dos artigos analisados, faz parte de uma série de conferências interessadas em abordagens inteligentes para transformar a *web* em uma máquina computacional global baseada em raciocínio e semântica. O objetivo do evento é prover um fórum para que pesquisadores apresentem contribuições de pesquisa, e é destinado a toda a comunidade da *web*, da indústria à pesquisa, dos desenvolvedores aos usuários e *designers*.

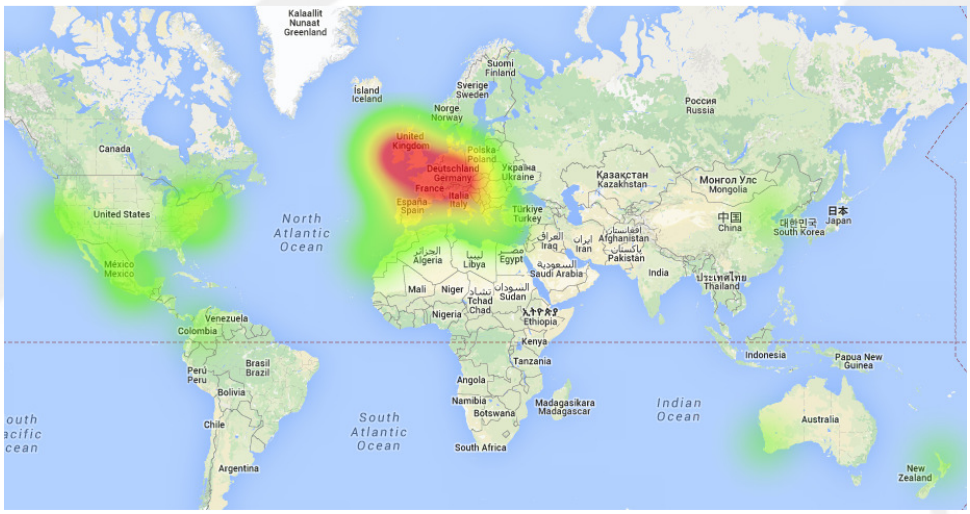
Também com dois artigos publicados, a "*International Semantic Web Conference (ISWC)*" é a principal conferência para pesquisas em tópicos da *Web* Semântica. O evento é realizado anualmente, sendo o sucessor do "*Semantic Web Working Symposium*" e, como particularidade, os dados sobre as conferências são disponibilizados para consulta em RDF.

ISBN 978-85-61115-09-8

3.5 Principais autores, instituições e países

Os 46 artigos analisados neste estudo foram escritos por 146 autores, de 55 instituições diferentes em 15 países. Os países com maior número de autores são: Irlanda, com 30; Alemanha, com 25; Espanha, com 18, Suíça, com 14; Reino Unido, com 11; e, Estados Unidos, com 10 autores. A Figura 2 apresenta um mapa com a distribuição das ocorrências das obras, segundo a localização das instituições às quais os autores estão vinculados. As áreas mais escuras indicam as maiores concentrações.

Figura 5- ***Mapa das ocorrências, segundo a localização das instituições de vínculo dos autores***



Fonte: elaborado pelos autores

Os sete autores com maior número de artigos são apresentados na tabela 4.

Tabela 4- ***Autores com maior número de artigos***

Autor	Instituição	Artigos	1º Autor
Edward Curry	DERI - National University of Ireland	3	2
Wassim Derguech	DERI - National University of Ireland	3	0
Manfred Hauswirth	DERI - National University of Ireland	3	0
Maciej Zaremba	DERI - National University of Ireland	3	0
José María Álvarez	Universidad de Oviedo	2	2
Feng Gao	DERI - National University of Ireland	2	2
Mathias Thoma	SAP Research Switzerland	2	2

Fonte: elaborado pelos autores

Autor de três dos artigos selecionados, sendo primeiro autor em dois deles, Edward Curry é um pesquisador do DERI e lidera o *Green and Sustainable IT research group*. Seus projetos de pesquisa incluem estudos de sustentabilidade em TI, inteligência energética, gestão de informações semânticas e gestão colaborativa de dados. Edward tem mais de 60 artigos publicados em *journals*, livros e conferências internacionais.

Wassim Derguech, coautor em três artigos da base de documentos analisada, é pesquisador e estudante de doutorado no DERI. Suas áreas de interesse incluem modelagem de processos de negócio, gestão de variabilidade em processos de negócios, reuso em processos de negócios,sustentabilidade em TI, sistemas de gestão de energia, tomada de decisão. Derguech tem 18 publicações nos últimos cinco anos.

Manfred Hauwirth, coautor de três dos artigos avaliados, é atualmente diretor-gerente do *Fraunhofer Institute for Open Communication Systems – FOKUS* é professor de Sistemas

ISBN 978-85-61115-09-8

Distribuídos Abertos na TU Berlin e também professor pesquisador na *National University of Ireland*, onde foi vice-diretor geral do DERI por oito anos. Suas áreas de interesse envolvem, dentre outros temas, *Linked Data streams*, *sensor networks*, *internet of things*, sistemas *peer-to-peer*.

Também com três artigos, Maciej Zaremba é pesquisador do *Digital Enterprise Research Institute - DERI* na *Service Oriented Architectures Unit* e conta com mais de 40 publicações relacionadas à *Linked Data*, *Web Services* e *Cloud Systems*.

José María Álvarez, autor de dois artigos, foi professor assistente de 2008 a 2012 no Departamento de Ciência da Computação na Universidade de Oviedo, onde também trabalhou no grupo de pesquisa WESO e onde fez seu PhD. Álvarez é autor de mais de 40 publicações e trabalhos de pesquisa, tendo sido agraciado por diversas premiações. Atualmente é professor visitante na *Universidad Carlos III*, em Madrid, onde pesquisa sobre Engenharia de Sistemas e Interoperabilidade.

Autor de dois artigos, Feng Gao trabalhou com o *Digital Enterprise Research Institute – DERI* no período de 2009 a 2013 como estudante de doutorado. Seu trabalho esteve vinculado à *Service Oriented Architectures Unit*.

Também tendo dois artigos de sua autoria, Mathias Thomma é pesquisador na *SAP Research Switzwerland*, nos grupos *Products and Innovation* (P&I), *Plattform Architecture and Technology Solutions* (PA&TS), e atualmente está trabalhando na área de *Internet of Things/Cyber-Physical Systems and Human-Computer-Interaction*.

Os autores dos artigos constantes do corpo de documentos selecionado para análise estão vinculados a 55 instituições. A Tabela 5 apresenta as instituições mais representativas em termos de número de autores vinculados.

Tabela 5 - ***Número de autores por instituição***

Instituição	País	Autores
National University of Ireland	Irlanda	26
IBM T.J. Watson Research Center	EUA	06
SAP Research France	França	06
Institute of Information Systems	Áustria	05
SAP AG Germany	Alemanha	05

Fonte: elaborado pelos autores

A instituição mais atuante, com 26 autores, é a *National University of Ireland* através do *Digital Enterprise Research Institute* (DERI). Trata-se é um Centro de Ciência, Engenharia e Tecnologia, estabelecido em 2003 pela *Science Foundation Ireland*, reconhecido internacionalmente em pesquisas sobre Web Semântica, educação e transferência de tecnologia. O objetivo do DERI é tornar-se reconhecido como o principal instituto de pesquisa sobre a web, interligando tecnologias, informações e pessoas para o avanço dos negócios e o benefício da sociedade. Atualmente, conta com mais de 140 membros, tendo aproximadamente 1500 publicações de pesquisa.

Seis dos autores têm vínculo com o *IBM Thomas J. Watson Research Center*, que é a sede da divisão de pesquisas da IBM e promove melhorias em *hardware*, serviços, *software* e sistemas, bem como a matemática e as ciências que suportam a indústria de tecnologia da informação.

A *SAP Research* é a unidade global de pesquisas em tecnologia da SAP AG, e conta com uma rede de 21 localizações no mundo inteiro. Seis dos artigos analisados foram publicados por autores vinculados à divisão francesa, e cinco vinculados à divisão alemã. A divisão de pesquisa contriui para o portfólio de produtos da SAP, desenvolvendo pesquisas aplicadas em temas que são tendência em tecnologias da informação.

O *Institute of Information Systems*, também com cinco autores, é vinculado à Faculdade de Informática da *Vienna University of Technology*. O instituto desenvolve

ISBN 978-85-61115-09-8

pesquisas sobre representação, armazenamento, distribuição, proteção e manipulação de dados e conhecimento. Os grupos de pesquisa do instituto dividem-se em: Sistemas Distribuídos; Banco de Dados e Inteligência Artificial; Sistemas Baseados em Conhecimento; Métodos Formais e Engenharia de Sistemas; e, Computação Paralela.

3.6 Principais referências utilizadas

Os 46 artigos analisados utilizaram um total de 903 referências bibliográficas, o que resulta em uma média de aproximadamente 19 referências por artigo. A tabela 6 apresenta as referências citadas por três ou mais dos artigos analisados. As quatro referências mais citadas são descritas a seguir.

Tabela 6 - ***Referências mais citadas***

Referência	Ano	Autores	Citações
Linked Data – The Story So Far	2009	Christian Bizer, Tom Heath e Tim Berners-Lee	11
Linked Data – Design Issues	2006	Tim Berners-Lee	08
SPARQL Query Language for RDF	2008	Eric Prud'hommeaux	07
Linked Data: Evolving the Web into a Global Data Space	2011	Tom Heath e Christian Bizer	04

Fonte: elaborado pelos autores

O trabalho mais vezes citado é "*Linked Data – The Story So far*"(BIZER, HEATH; BERNERS-LEE, 2009), com 11 ocorrências, apresenta o conceito e os princípios técnicos de *Linked Data* e o progresso na publicação de *Linked Data* na *Web*, revisando aplicações que foram desenvolvidas e traçando uma agenda para os próximos passos da comunidade *Linked Data*.

Citado oito vezes, "*Linked Data – Design Issues*" (BERNERS-LEE, 2006) é o artigo onde o autor apresenta os princípios, ou "as quatro regras", de *Linked Data*, e discute soluções, detalhes de implementação e fatores que afetam as escolhas sobre a forma de se publicar dados na *web*.

Com sete citações, "*SPARQL Query Language for RDF*" (Prud'hommeaux, 2008) é uma recomendação da W3C que descreve a linguagem SPARQL, que permite realizar consultas sobre dados em RDF. O documento é uma fonte de referência técnica para desenvolvedores, e apresenta características, convenções, sintaxe e exemplos de utilização.

O quarto artigo mais citado, com quatro ocorrências, é "*Linked Data: Evolving the Web into a Global Data Space*" (HEATH; BIZER, 2011).Trata-se de um livro que apresenta uma visão geral dos princípios de *Linked Data* e da *Web* de Dados que emergiu graças à aplicação destes princípios. Os autores discutem padrões para a publicação de *Linked Data*, descrevem aplicações desenvolvidas e examinam suas arquiteturas.

4. Considerações Finais

A pesquisa realizada nas três bases de dados: Scopus, Web of Science e *IEEE Xplore Digital Library,* resultou na análise de 46 artigos sobre o termo, um método considerado ainda novo, de padronização da criação da informação e sua disponibilidade e recuperabilidade pela própria organização. Acredita-se que este padrão apresenta-se como uma forma mais efetiva de realizar esse tipo de atividade.

Dos documentos utilizados para análise, apenas dois autores trouxeram a definição do termo *Enterprise Linked Data,* e um trouxe cenários possíveis para as empresas utilizarem os dados ligados. Percebe-se com essa pesquisa que ainda existe muito sobre o termo *Enterprise Linked Data* para se explorar, tanto na academia como pelas organizações.

ISBN 978-85-61115-09-8

Quanto às publicações, as maiores incidências de publicação aconteceram nos anos de 2011 e 2013 somando 12 documentos publicados em cada ano.

A fonte que mais apareceu citada nos artigos foi o livro "Linking Enterprise Data" de Danid Wood, publicado em 2010. O Journal Advanced Engineering Informatics, da Elsevier e as Conferências Internacionais: HCISS, I-SEMANTICS, WIMS e ISWC também estão na lista das fontes que mais apareceram nessa pesquisa.

Os documentos analisados contam com 146 autores de 55 instituições diferentes em 15 países. Os países com maior índice de autores são: Irlanda, Alemanha, Espanha, Suíça, Reino Unido e Estados Unidos. Os sete autores que mais apareceram nessa análise foram: Edward Curry, Wassim Derguech, Manfred Hauswirth, Maciej Zaremba, José María Álvarez, Feng Gao e Mathias Thoma.

A instituição mais atuante é a National University of Ireland, seguida pela IBM T. J. Watson Research Center, SAP Research France, Institute of Information Systems e SAP AG Germany. Referências bibliográficas utilizadas pelos 46 documentos totalizaram 903 fontes. A mais citada foi: Linked Data – The Story So far de Christian Bizer, Tom Heath e Tim Berners-Lee de 2009. Seguida pelas obras: Linked Data – Design Issues (2006) de Tim Berners-Lee, SPARQL Query Language for RDF (2008) de Eric Prud'hommeaux e Linked Data: Evolving the Web into a Global Data Space (2011) de Tom Heath e Christiam Bizer.

Conclui-se que os dados apresentados sobre o tema *Enterprise Linked Data*, como as referências levantadas, autores, instituições e as outras informações são de grande relevância para pesquisas futuras.

Referências

ABBAS, S.; OJO, A. Applying Design Patterns in URI Strategies -- Naming in Linked Geospatial Data Infrastructure. In: SYSTEM SCIENCES (HICSS), 2014 47TH HAWAII INTERNATIONAL CONFERENCE ON, 2014. 2014. p. 2094-2103.

ALLEMANG, D. Semantic web and the linked data enterprise. In: (Org.). **Linking Enterprise Data**: Springer US, 2010. p. 3-23. ISBN 9781441976642 (ISBN).

ALOR-HERNÁNDEZ, G. *et al.* BROSEMWEB: A brokerage service for e-Procurement using Semantic Web Technologies. **Computers in Industry**,v. 65, n. 5, p. 828-840, 2014.

ÁLVAREZ, J. M. *et al.* Query expansion methods and performance evaluation for reusing linking open data of the European public procurement notices. In: 14TH CONFERENCE OF THE SPANISH ASSOCIATION FOR ARTIFICIAL INTELLIGENCE, CAEPIA 2011, 2011. La Laguna. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. La Laguna, 2011. p. 494-503.

ALVAREZ, J. M. *et al.*Towards a pan-european e-procurement platform to aggregate, publish and search public procurement notices powered by linked open data: The moldeas approach. **International Journal of Software Engineering and Knowledge Engineering**,v. 22, n. 3, p. 365-383, 2012.

ARAÚJO, C. A. **Bibliometria: evolução histórica e questões atuais**. Em questão, Porto Alegre, v. 12, n. 1, p. 11-32, jan./jun. 2006.

ASSAF, A. *et al.* RUBIX: A framework for improving data integration with linked data. In: 1ST INTERNATIONAL WORKSHOP ON OPEN DATA, WOD 2012, 2012. Nantes. **ACM International Conference Proceeding Series**. Nantes, 2012. p. 13-21.

BERNERS-LEE, Tim. **Linked Data: Design Issues** 2006.

ISBN 978-85-61115-09-8

BHIRI, S.; DERGUECH, W.; ZAREMBA, M. Modelling capabilities as attribute-featured entities. In: 8TH INTERNATIONAL CONFERENCE ON WEB INFORMATION SYSTEMS AND TECHNOLOGIES, WEBIST 2012, 2013. Porto. **Lecture Notes in Business Information Processing**. Porto, 2013. p. 70-85.

BIANCHINI, D.; DE ANTONELLIS, V.; MELCHIORI, M. A linked data perspective for collaboration in mashup development. In: 24TH INTERNATIONAL WORKSHOP ON DATABASE AND EXPERT SYSTEMS APPLICATIONS, DEXA 2013, 2013. Prague. **Proceedings - International Workshop on Database and Expert Systems Applications, DEXA**. Prague, 2013. p. 128-132.

BIZER, Chris; CYGANIAK, Richard; HEATH, Tom. **How to publish Linked Data on the Web** 2007.

BIZER, Chris; HEATH, Tom; BERNERS-LEE, Tim. Linked data - The story so far. **International Journal on Semantic Web and Information Systems**,v. 5, n. 3, p. 22, 2009.

BIZER, Christian *et al.* DBPedia - A Crystalization Point for the Web of Data. 2009.

BRAMBILLA, S. D. S.; STUMPF, I. R. C. Produção científica da UFRGS representada na Web of Science Scientific production of UFRGS INDEXED at Web of Science (2000-. **Perspectivas em Ciência da Informação**,v. 17, n. 3, p. 17, 2012.

CHENG, L. *et al.*Runtime characterization of triple stores. In: 15TH IEEE INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE AND ENGINEERING, CSE 2012 AND 10TH IEEE/IFIP INTERNATIONAL CONFERENCE ON EMBEDDED AND UBIQUITOUS COMPUTING, EUC 2012, 2012. Paphos. **Proceedings - 15th IEEE International Conference on Computational Science and Engineering, CSE 2012 and 10th IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, EUC 2012**. Paphos, 2012. p. 66-73.

CHRISTIDIS, K.; MENTZAS, G.; APOSTOLOU, D. Supercharging enterprise 2.0. **IT Professional**,v. 13, n. 4, p. 29-35, 2011.

CORRY, E. *et al.* Using semantic web technologies to access soft AEC data. **Advanced Engineering Informatics**, 2014.

CURRY, E. System of systems information interoperability using a linked dataspace. In: 2012 7TH INTERNATIONAL CONFERENCE ON SYSTEM OF SYSTEMS ENGINEERING, SOSE 2012, 2012. Genova. **Proceedings - 2012 7th International Conference on System of Systems Engineering, SoSE 2012**. Genova, 2012. p. 101-106.

CURRY, E. *et al.* Linking building data in the cloud: Integrating cross-domain building data using linked data. **Advanced Engineering Informatics**,v. 27, n. 2, p. 206-219, 2013.

FENG, Gao *et al.* Extending BPMN 2.0 with Sensor and Smart Device Business Functions. In: ENABLING TECHNOLOGIES: INFRASTRUCTURE FOR COLLABORATIVE ENTERPRISES (WETICE), 2011 20TH IEEE INTERNATIONAL WORKSHOPS ON, 2011. 2011. p. 297-302.

FOGAROLLI, A.; BOUQUET, P. Linking data for public administrations. In: 6TH INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, I-SEMANTICS '10, 2010. Graz. **ACM International Conference Proceeding Series**. Graz, 2010. p.

GAO, F.; DERGUECH, W.; ZAREMBA, M. Extending BPMN 2.0 to enable links between process models and ARIS views modeled with linked data. In: 4TH INTERNATIONAL CONFERENCE ON BUSINESS PROCESS AND SERVICES COMPUTING, BPSC 2011, IN CONJUNCTION WITH THE 14TH INTERNATIONAL CONFERENCE ON BUSINESS INFORMATION SYSTEMS, BIS 2011, 2011. Poznan. **Lecture Notes in Business Information Processing**. Poznan, 2011. p. 41-52.

GRAUBE, M. *et al.* Linked data as integrating technology for industrial data. In: 2011 INTERNATIONAL CONFERENCE ON NETWORK-BASED INFORMATION SYSTEMS, NBIS 2011, 2011. Tirana. **Proceedings - 2011 International Conference on Network-Based Information Systems, NBiS 2011**. Tirana, 2011. p. 162-167.

ISBN 978-85-61115-09-8

HALB, W. *et al.* Towards a commercial adoption of linked open data for online content providers. In: 6TH INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, I-SEMANTICS '10, 2010. Graz. **ACM International Conference Proceeding Series**. Graz, 2010. p.

HARRIS, K. Selling and building linked data: Drive value and gain momentum. In: (Org.). **Linking Enterprise Data**: Springer US, 2010. p. 65-76. ISBN 9781441976642 (ISBN).

HARRIS, S.; ILUBE, T.; TUFFIELD, M. Enterprise linked data as core business infrastructure. In: (Org.). **Linking Enterprise Data**: Springer US, 2010. p. 203-219. ISBN 9781441976642 (ISBN).

HASSANZADEH, O. *et al.* Helix: Online enterprise data analytics. In: 20TH INTERNATIONAL CONFERENCE COMPANION ON WORLD WIDE WEB, WWW 2011, 2011. Hyderabad. **Proceedings of the 20th International Conference Companion on World Wide Web, WWW 2011**. Hyderabad, 2011. p. 225-228.

HEATH, Tom; BIZER, Christian. **Linked Data: Evolving the Web into a Global Data Space** Morgan & Claypool**:** 136 p. 2011.

HOFFMANN, M. *et al.* Multi-dimensional production planning using a vertical data integration approach: A contribution to modular factory design. In: 2013 10TH INTERNATIONAL CONFERENCE AND EXPO ON EMERGING TECHNOLOGIES FOR A SMARTER WORLD, CEWIT 2013, 2013. Melville, NY. **2013 10th International Conference and Expo on Emerging Technologies for a Smarter World, CEWIT 2013**. Melville, NY: IEEE Computer Society, 2013. p.

HOVER, K. M.; MUHLHAUSER, M. Integrating distributed discussions in web 2.0 applications and their integration with lecture recordings. In: 15TH IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA, ISM 2013, 2013. Anaheim, CA. **Proceedings - 2013 IEEE International Symposium on Multimedia, ISM 2013**. Anaheim, CA, 2013. p. 486-491.

HU, B.; SVENSSON, G. A case study of linked enterprise data. In: 9TH INTERNATIONAL SEMANTIC WEB CONFERENCE, ISWC 2010, 2010. Shanghai. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. Shanghai, 2010. p. 129-144.

HYLAND, B. Preparing for a linked data enterprise. In: (Org.). **Linking Enterprise Data**: Springer US, 2010. p. 51-64. ISBN 9781441976642 (ISBN).

KIRRANE, S. DC proposal: Knowledge based access control policy specification and enforcement. In: 10TH INTERNATIONAL SEMANTIC WEB CONFERENCE, ISWC 2011, 2011. Bonn. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. Bonn, 2011. p. 293-300.

LANTHALER, M.; GUTL, C. A semantic description language for RESTful Data Services to combat Semaphobia. In: 5TH IEEE INTERNATIONAL CONFERENCE ON DIGITAL ECOSYSTEMS AND TECHNOLOGIES, DEST 2011, 2011. Daejeon. **IEEE International Conference on Digital Ecosystems and Technologies**. Daejeon, 2011. p. 47-53.

NASEER, A.; LAERA, L.; MATSUTSUKA, T. Enterprise BigGraph. In: SYSTEM SCIENCES (HICSS), 2013 46TH HAWAII INTERNATIONAL CONFERENCE ON, 2013. 2013. p. 1005-1014.

OMITOLA, T. *et al.* Linking social, open, and enterprise data. In: 4TH INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, MINING AND SEMANTICS, WIMS 2014, 2014. Thessaloniki. **ACM International Conference Proceeding Series**. Thessaloniki: Association for Computing Machinery, 2014. p.

ORTIZ, P. *et al.* Enhanced multi-domain access control for secure mobile collaboration through Linked Data cloud in manufacturing. In: WORLD OF WIRELESS, MOBILE AND MULTIMEDIA NETWORKS (WOWMOM), 2013 IEEE 14TH INTERNATIONAL SYMPOSIUM AND WORKSHOPS ON A, 2013. 2013. p. 1-9.

ISBN 978-85-61115-09-8

PAGANO, A. EUD in enterprise open source learning environments. In: 3RD INTERNATIONAL SYMPOSIUM ON END-USER DEVELOPMENT, IS-EUD 2011, 2011. Torre Canne. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. Torre Canne, 2011. p. 363-366.

PASSANT, A. *et al.* Enhancing enterprise 2.0 ecosystems using semantic web and linked data technologies: The SemSLATES approach. In: (Org.). **Linking Enterprise Data**: Springer US, 2010. p. 79-102. ISBN 9781441976642 (ISBN).

PIZZANI, L.; SILVA, R. C.; HAYASHI, M. C. P. I. **Bases de dados e bibliometria**: a presença da educação especial na base Medline. Revista brasileira de biblioteconomia e documentação, v.4, n.1, p. 68-85, jan./jun. 2008.

ROHDE, M. E.; SUNDARAM, D. Knowledge Composition: Theory, Architecture and Implementation. In: INFORMATION, PROCESS, AND KNOWLEDGE MANAGEMENT, 2010. EKNOW '10. SECOND INTERNATIONAL CONFERENCE ON, 2010. 2010. p. 80-85.

RUSITSCHKA, S. *et al.* Adaptive middleware for real-time prescriptive analytics in large scale power systems. In: INDUSTRIAL TRACK OF THE 13TH ACM/IFIP/USENIX INTERNATIONAL MIDDLEWARE CONFERENCE, MIDDLEWARE INDUSTRY 2013, 2013. Beijing. **Proceedings of the Industrial Track of the 13th ACM/IFIP/USENIX International Middleware Conference, Middleware Industry 2013**. Beijing: Association for Computing Machinery, 2013. p.

SACCO, O.; PASSANT, A.; DECKER, S. An Access Control Framework for the Web of Data. In: TRUST, SECURITY AND PRIVACY IN COMPUTING AND COMMUNICATIONS (TRUSTCOM), 2011 IEEE 10TH INTERNATIONAL CONFERENCE ON, 2011. 2011. p. 456-463.

SALMEN, A. *et al.* ComVantage: Mobile enterprise collaboration reference framework and enablers for future internet information interoperability. In: 2013 ANNUAL FUTURE INTERNET ASSEMBLY, FIA 2013, 2013. Dublin. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. Dublin: Springer Verlag, 2013. p. 220-232.

SERRANO, M. *et al.* A Self-Organizing Architecture for Cloud by Means of Infrastructure Performance and Event Data. In: CLOUD COMPUTING TECHNOLOGY AND SCIENCE (CLOUDCOM), 2013 IEEE 5TH INTERNATIONAL CONFERENCE ON, 2013. 2013. p. 481-486.

SERVANT, F. P. Linking enterprise data. In: WWW 2008 WORKSHOP ON LINKED DATA ON THE WEB, LDOW 2008, 2008. Beijing. **CEUR Workshop Proceedings**. Beijing, 2008. p.

SHAOPENG, He; JIANHUI, Li; ZHIHONG, Shen. F2R: Publishing file systems as Linked Data. In: FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY (FSKD), 2013 10TH INTERNATIONAL CONFERENCE ON, 2013. 2013. p. 767-772.

SLEIMAN, H. A.; RIVERO, C. R.; CORCHUELO, R. On a proposal to integrate web sources using semantic-web technologies. In: 2011 7TH INTERNATIONAL CONFERENCE ON NEXT GENERATION WEB SERVICES PRACTICES, NWESP 2011, 2011. Salamanca. **Proceedings of the 2011 7th International Conference on Next Generation Web Services Practices, NWeSP 2011**. Salamanca, 2011. p. 326-331.

THOMA, M. *et al.* Linked services for M2M communication with Enterprise IT systems. In: 2013 9TH INTERNATIONAL WIRELESS COMMUNICATIONS AND MOBILE COMPUTING CONFERENCE, IWCMC 2013, 2013. Cagliari, Sardinia. **2013 9th International Wireless Communications and Mobile Computing Conference, IWCMC 2013**. Cagliari, Sardinia, 2013. p. 1212-1216.

THOMA, M.; SPERNER, K.; BRAUN, T. Service descriptions and linked data for integrating WSNs into enterprise IT. In: 2012 3RD INTERNATIONAL WORKSHOP ON SOFTWARE ENGINEERING FOR SENSOR NETWORK APPLICATIONS, SESENA 2012, 2012. Zurich. **2012 3rd International Workshop on Software Engineering for Sensor Network Applications, SESENA 2012 - Proceedings**. Zurich, 2012. p. 43-48.

ISBN 978-85-61115-09-8

TUÁN, A. L. *et al.* Global sensor modeling and constrained application methods enabling cloud-based open space smart services. In: 9TH IEEE INTERNATIONAL CONFERENCE ON UBIQUITOUS INTELLIGENCE AND COMPUTING, UIC 2012 AND 9TH IEEE INTERNATIONAL CONFERENCE ON AUTONOMIC AND TRUSTED COMPUTING, ATC 2012, 2012. Fukuoka. **Proceedings - IEEE 9th International Conference on Ubiquitous Intelligence and Computing and IEEE 9th International Conference on Autonomic and Trusted Computing, UIC-ATC 2012**. Fukuoka, 2012. p. 196-203.

UMBRICH, J. *et al.* Linked Data and Live Querying for Enabling Support Platforms for Web Dataspaces. In: DATA ENGINEERING WORKSHOPS (ICDEW), 2012 IEEE 28TH INTERNATIONAL CONFERENCE ON, 2012. 2012. p. 23-28.

VANZ, S. A. de S.; CAREGNATO, S. E. **Estudos de citação**: uma ferramenta para entender a comunicação científica. Em Questão, Porto Alegre, v. 9, n. 2, p. 295-307, jul./dez. 2003.

VILCHES-BLÁZQUEZ, Luis M. *et al.* An Approach to Publish Spatial Data on the Web: The GeoLinked Data Case. In: WORKSHOP ON LINKED SPATIOTEMPORAL DATA 2010 IN CONJUNCTION WITH THE 6TH INTERNATIONAL CONFERENCE ON GEOGRAPHIC INFORMATION SCIENCE , GISCIENCE 2010. Zurich, Suiça, 2010. p.

WEICHSELBRAUN, A.; STREIFF, D.; SCHARL, A. Linked enterprise data for fine grained named entity linking and web intelligence. In: 4TH INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, MINING AND SEMANTICS, WIMS 2014, 2014. Thessaloniki. **ACM International Conference Proceeding Series**. Thessaloniki: Association for Computing Machinery, 2014. p.

WESTERSKI, A.; IGLESIAS, C. A. Exploiting structured linked data in enterprise knowledge management systems: An idea management case study. In: 15TH IEEE INTERNATIONAL EDOC ENTERPRISE COMPUTING CONFERENCE WORKSHOPS, EDOCW 2011, 2011. Helsinki. **Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC**. Helsinki, 2011. p. 395-403.

WOOD, D. **Linking enterprise data**. Springer US, 2010. 1-291 p. ISBN 9781441976642 (ISBN).

WURZER, J. Building a bridge between information and process management. In: 9TH INTERNATIONAL CONFERENCE ON BUSINESS PROCESS MANAGEMENT, BPM 2011P, 2012. Clermont-Ferrand. **Lecture Notes in Business Information Processing**. Clermont-Ferrand, 2012. p. 318-319.

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

INFRAESTRUTURA DE INFORMAÇÃO PARA TRANSPARÊNCIA E APOIO À GESTÃO DE AÇÕES GOVERNAMENTAIS EM SANTA CATARINA

Maicon Guzzon Lima
maicon.guzzon@gmail.com

Cristiano Cortez da Rocha
cristiano.darocha@gmail.com

Guilherme Kraus dos Santos
gkraus@sef.sc.gov.br

Fábio José do Amaral
amaral@ciasc.sc.gov.br

Resumo
A administração pública brasileira historicamente não possui a preocupação de acompanhar e avaliar programas e políticas públicas nem de apresentar de forma transparente para a sociedade as informações referentes. Entretanto, é inevitável a utilização de um processo de gestão eficiente, com o objetivo de subsidiar o planejamento e a formulação das intervenções governamentais. Neste sentido, este trabalho relata a experiência de aplicação de um processo de acompanhamento físico e financeiro das ações governamentais nas Audiências Públicas Regionalizadas do Governo do Estado de Santa Catarina. Além disso, a partir deste estudo de caso realizado, este trabalho propõe uma nova infraestrutura de informação para atender as exigências estabelecidas para dados abertos governamentais e atender as necessidades analíticas da alta gestão governamental.

Palavras-chave: Acompanhamento de Ações Governamentais. Transparência Pública. Dados Abertos Governamentais.

Abstract
The brazilian public management is not used to monitor, to assess public policies and programs, and to present them in a transparent way to the society. However, an efficient management process is highly required in order to support the planning and creation of governmental actions. In this context, the present work reports the experience of using a physical and financial monitoring process of the governmental actions in the Public Audiences of the Government of the Santa Catarina State. Moreover, this work proposes a new information infrastructure to meet the open government data's principles and to comply with the analytical needs of the governmental management.

Key Words: Governmental Actions Monitoring. Public Transparency. Open Government Data.

1 Introdução

Com o surgimento da Lei 12.527/12, o Brasil dá um grande passo no que diz respeito à transparência pública. É através dessa lei que torna-se direito de qualquer cidadão o acesso às informações produzidas ou custodiadas pelo poder público, como folha de pagamento de

ISBN 978-85-61115-09-8

servidores, investimentos em educação, saúde, segurança, gastos com manutenção, terceirização de serviços, entre outros.

Além disso, como se não bastasse a importância do acesso a estas informações por parte da população, tanto para cumprimento de lei como também para uma melhora significativa na transparência pública, elas também devem estar acessíveis aos diversos órgãos do governo. Só assim seria possível realizar o controle e o monitoramento da execução orçamentária de uma forma mais eficaz por parte dos gestores públicos. Como complemento a isso, vale citar também que a existência de uma política para lidar com a informação produzida já é prevista em lei. De acordo com a Constituição Estadual de Santa Catarina, os poderes Executivo, Legislativo e Judiciário devem possuir uma ferramenta integrada de avaliação e monitoramento das metas previstas no plano plurianual bem como do andamento da execução dos programas de governo e orçamentos do Estado.

Como consequência da necessidade governamental de transmitir informações, tanto internamente, de um órgão para outro, quanto para a sociedade, torna-se de fundamental importância a criação de uma infraestrutura de informação que dê suporte a qualquer aplicação que tenha o objetivo de utilizar as informações de forma rápida e consistente ou que necessite trabalhar com diferentes tipos e combinações de filtros Além disso, é importante permitir uma fácil manutenibilidade, dada novas necessidades dos gestores ou eventuais mudanças na legislação, bem como possibilitar o manuseio da informação na forma G2G (governo para governo), G2B (governo para empesa) ou G2C (governo para cidadão).

Portanto o presente trabalho propõe uma nova infraestrutura de informação para apoio ao acompanhamento físico e financeiro de todas as ações planejadas e realizadas pelo Governo do Estado de Santa Catarina. O trabalho deve-se a finalidade de suportar o processo de tomada de decisão da alta gestão, através da análise dos diversos indicadores que podem ser providos e construídos pela infraestrutura proposta.

O trabalho é apresentado da seguinte forma: na seção 2 são apresentados aspectos teóricos sobre o acompanhamento das ações governamentais, juntamente com a motivação para o projeto dessa nova abordagem. Na seção 3 são apresentados os procedimentos metodológicos utilizados, juntamente com estudo de caso realizado. Na seção 4 é apresentada a infraestrutura proposta para lidar com as necessidades apontadas no estudo de caso e, por fim, na seção 5 são apresentadas as considerações finais, juntamente com algumas sugestões para trabalhos futuros.

2 Fundamentos Teóricos

Essa seção apresenta os aspectos teóricos relacionados ao monitoramento e avaliação no ciclo do planejamento de ações governamentais.

2.1 Processo Orçamentário

Antes da promulgação da Constituição Federal de 1988, não havia no Brasil um processo bem definido para o planejamento e a execução do orçamento público, ficando a cargo de cada um dos entes da Federação a competência legislativa para tanto. Dentre as inúmeras leis surgidas naquele período, destacam-se a lei 4.320/64, conhecida como Quadro de Recursos e de Aplicação de Capital (QRAC) e a lei do Orçamento Plurianual de Investimentos (OPI), introduzida no país através da Constituição de 1967, que modificou ligeiramente a anterior ao criar a necessidade de aprovar as receitas e as despesas através de lei elaborada pelo poder Executivo e não mais por decreto como era feito no QRAC.

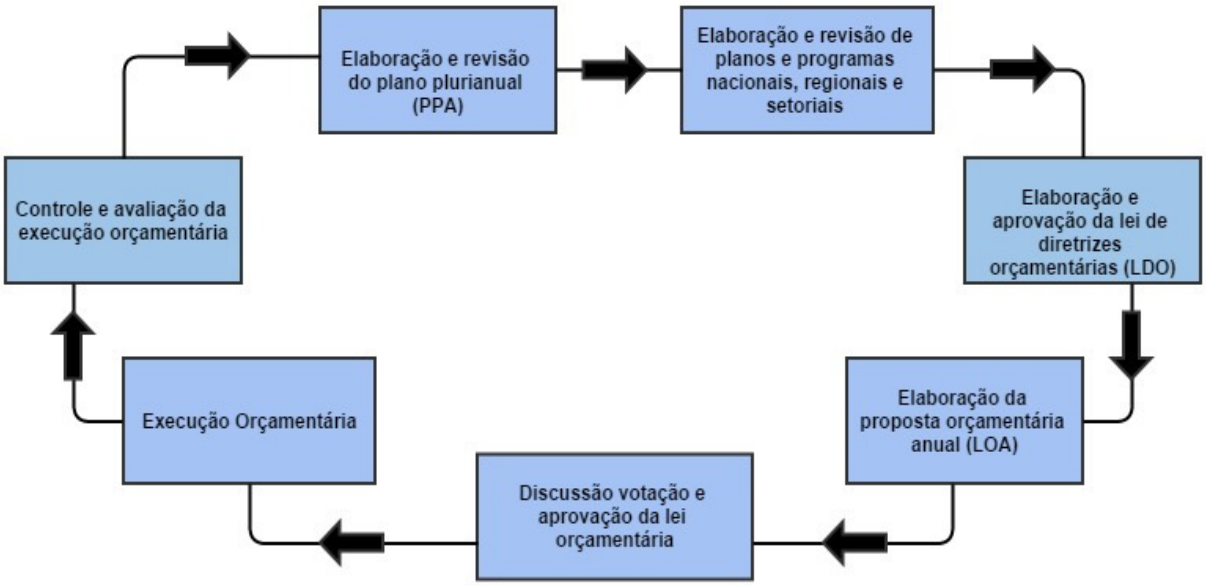
Com o intuito de eliminar problemas como a burocracia causada pela necessidade de revisão e atualização anual do OPI, afinal ele não possuía caráter plurianual, assim como pelo

ISBN 978-85-61115-09-8

fato deste considerar apenas despesas de capital e não despesas correntes, foi introduzido no país pela Constituição de 1988 diretrizes inovadoras com a criação de novos instrumentos como o Plano Plurianual (PPA), Lei de Diretrizes Orçamentárias (LDO) e a Lei Orçamentária Anual (LOA) que juntas conseguem ligar planos de médio prazo com orçamentos anuais, incluindo também todas as receitas e despesas no processo orçamentário.

Na Figura 1 é mostrado de forma estática como se dá o ciclo do processo integrado de planejamento e orçamento. Percebe-se a existência de uma etapa denominada "*controle e avaliação da execução orçamentária",* em que, para um cumprimento eficiente daquilo que efetivamente deve ser feito nesta etapa (controle e avaliação), é indispensável uma ferramenta que gerencie e forneça as informações desejadas de uma forma clara e em tempo hábil para os gestores do PPA.

Figura 6 - ***Ciclo Orçamentário***



Fonte: Adaptado de Giacomoni (2010)

2.1.1 Plano Plurianual

O Plano Plurianual, conforme apresenta Giacomoni (2010), é uma das principais novidades do novo marco constitucional, representando a síntese dos esforços de planejamento de toda a administração pública, orientando a elaboração dos demais planos e programas de governo, assim como do próprio orçamento anual. Em outras palavras, representa um resumo de todo o planejamento da administração pública.

De acordo com Chafun (2001) o PPA é entendido como um instrumento para evidenciar programas de trabalho de governo no qual se enfatizam políticas, diretrizes e ações programadas no longo prazo bem como os respectivos objetivos a serem alcançados, devidamente quantificados fisicamente.

Quanto à vigência do PPA, a Constituição Federal de 1988 estabelece que o plano cobrirá o período compreendido entre o início do segundo ano do mandato e o final do primeiro exercício do mandato subsequente. Portanto, um PPA constitui-se a partir do programa do governo que assumiu no ano anterior, e estabelece um novo arranjo de planejamento, orçamento e gerenciamento de programas. Esta vigência dessincronizada do mandato eleitoral tem o objetivo de evitar a descontinuidade das políticas públicas e de fornecer tempo ao novo governante, durante o primeiro ano de exercício, para analisar e

ISBN 978-85-61115-09-8

estruturar seu plano e suas diretrizes estratégicas para ospróximos quatro anos (GIACOMONI, 2010).

2.1.2 Lei de Diretrizes Orçamentárias – LDO

A LDO é a lei que determina as metas e prioridades da administração pública, incluindo as despesas de capital para o exercício financeiro subsequente. Podem ser consideradas despesas de capital qualquer compra de bens realizada, tais como aquisição de veículos, construção de hospitais e escolas, por exemplo. De acordo com Giacomoni (2010), a Constituição incumbe a LDO disciplinar outros importantes assuntos, cuja definição antecipada representa importante apoio na preparação do projeto de lei orçamentária. Um exemplo são as autorizações para a concessão de qualquer vantagem ou aumento de remuneração, para a criação de cargos, empregos e funções ou alterações de estrutura e carreiras, bem como para a admissão ou contratação de pessoal, a qualquer título, pelos órgãos e entidades da administração direta e indireta, ressalvadas as empresas públicas e sociedades de economia mista.

Em Santa Catarina, dentre as disposições desta Lei, estão a definição das metas e das prioridades da Administração Pública Estadual, a organização e estrutura dos orçamentos, a limitação do percentual de despesas dos poderes Legislativo, Judiciário, Ministério Público e da Universidade do Estado de Santa Catarina (UDESC), além de monitorar e avaliar os instrumentos de planejamento.

2.1.3 Lei Orçamentária Anual - LOA

A LOA é o ato pelo qual o Poder Legislativo prevê e autoriza, em pormenor, a administração pública a realizar as despesas necessárias ao funcionamento dos serviços públicos, e a outros fins necessários às políticas públicas estabelecidas. De acordo com a Constituição Federal, a LOA é encaminhada até quatro meses antes do encerramento do exercício financeiro e devolvido para sanção até o encerramento da sessão legislativa. Já a Constituição Estadual de Santa Catarina de 1989, determina que a LOA deverá ser encaminhada até três meses antes do encerramento do exercício financeiro

2.2 Legislação: Obrigatoriedade de avaliar, monitorar e disponibilizar

A obrigatoriedade de avaliar e monitorar o orçamento não só é importante como também prevista em lei. Pode-se perceber abaixo alguns trechos da legislação que denotam tal obrigatoriedade:

> CE de 89 - Art. 62.
> Os Poderes Legislativo, Executivo e Judiciário manterão, de forma integrada, sistema de controle interno com a finalidade de:
> I - avaliar o cumprimento das metas previstas no plano plurianual, a execução dos programas de governo e dos orçamentos do Estado...
> Lei nº 15.722 de 2011 - Art. 12.
> Os órgãos do Poder Executivo, abrangendo seus fundos, autarquias, fundações, empresas públicas e sociedades de economia mista, pertencentes aos Orçamentos Fiscal, da Seguridade Social e de Investimento, responsáveis por programas e sub ações nos termos do Anexo Único desta Lei, deverão manter atualizadas, durante cada exercício financeiro, as informações referentes à execução física das sub ações sob sua responsabilidade, na forma estabelecida pelo órgão central do Sistema de Planejamento e Orçamento.
>
> DECRETO Nº 1.324, de 21 de dezembro de 2012: Este decreto estabeleceu em lei o processo de acompanhamento físico e financeiro e de avaliação do Plano

ISBN 978-85-61115-09-8

> Plurianual (PPA), onde passa a ser de obrigação do estado a divulgação de informações referentes aos resultados alcançados pela ação governamental. Salienta também em diversos trechos, a obrigação do monitoramento do PPA, do orçamento em si e o acompanhamento das metas físicas e financeiras. Isso pode ser visto no Art. 8: "monitorar a realização das metas física e financeira relativas ao objeto de execução, vinculado a uma subação e a um programa do PPA".

Entretanto, o modelo atual do processo não é aceito como unanimidade entre os especialistas. De acordo com Waterson (apud Giacomoni, 2010), é um dosespecialistas que defende a inviabilidade do modelo para países em desenvolvimento. A Tabela 1 apresenta algumas diferenças entre o modelo de planejamento atual e o defendido por Waterson.

Tabela 3- Quadro comparativo entre o modelo convencional e o centrado nos problemas

Convencional	Modelo de Waterson: centrado nos problemas
Estabelecimento de objetivos	Determinação dos problemas sociais básicos que devem ser resolvidos
Fixação de metas	Adaptação dos recursos disponíveis a esses problemas
Formulação da estratégia para alcançar as metas	Seleção de projetos e políticas que contribuem para a resolução de problemas
Seleção de políticas e projetos	Formulação de estratégia para resolver problemas
Solução de problemas sociais básicos	Seleção de projetos gerais conforme problemas sociais que devem ser resolvidos

Fonte: Adaptado de Giacomoni (2010)

Esse modelo de processo defende que a instabilidade política, a incerteza econômica, a ineficiência administrativa, a ausência de dados além das deficiências técnicas dos países em desenvolvimento, torna o modelo atual ineficaz para planos de médio e longo prazo.

Outro especialista contra o modelo atual é Wildavsky (apud Giacomoni, 2010) o qual defende a ideia que "o principal fator determinante do tamanho e do conteúdo do orçamento deste ano é o orçamento do ano passado", fato que permite pouca flexibilidade na elaboração de cada orçamento.

2.1.3 Procedimentos metodológicos

Enquadrar o estudo em uma metodologia específica pressupõe um esforço prévio de delineação da própria natureza da pesquisa, bem como de seu objetivo e sua abordagem do problema. Além disso, é necessário definir os procedimentos técnicos a serem adotados e o processo de coleta de dados da pesquisa. Quanto à natureza da pesquisa, trata-se de uma pesquisa aplicada. Segundo Barros e Lehfeld (2000, p.78), a pesquisa aplicada tem como finalidade e motivação gerar conhecimento para aplicação de seus resultados tendo como meta contribuir para a prática, visando à resolução do problema identificado na realidade. Salienta Appolinário (2004) afirma que resolver necessidades ou problemas concretos é o principal objeto de uma pesquisa aplicada. No que tange ao seu objetivo, a pesquisa é considerada descritiva pois, de acordo com Gil (1991), estas pesquisas visam descrever características de um determinado fenômeno.

O procedimento técnico adotado foi o de uma pesquisa-ação, pois a pesquisa foi concebida e realizada em uma estreita associação da ação e da resolução de um problema encontrado no Governo do Estado de Santa Catarina. De acordo com Gil (1991) os pesquisadores estão envolvidos de modo cooperativo e participativo, assim fazendo parte da mudança.

ISBN 978-85-61115-09-8

Por conta da abordagem do problema adotada, pode-se classificar o presente estudo como uma pesquisa tecnológica. Isso depreende-se do fato de o estudo objetivar a solução de um problema já existente no campo prático, e ter a sua aplicação no cenário real como o seu eixo central. Dessa forma, o avanço tecnológico proporcionado pela utilização efetiva dos resultados da pesquisa - materializados no sistema que se tornou seu resultado - causou um impacto direto e de fácil percepção no ambiente em que foi inserido.

A coleta de dados foi realizada por meio de dados primários e secundários. Segundo Mattar (2005, p. 159), são considerados dados primários aqueles nunca antes coletados. Para este estudo os dados primários foram os levantados por meio da observação sistemática do processo orçamentário junto à Diretoria de Planejamento Orçamentário da Secretaria de Estado da Fazenda de Santa Catarina, responsável pelo sistema administrativo de planejamento e orçamento. Já os dados secundários são aqueles já coletados, ordenados, tabulados e até mesmo analisados com a finalidade de atender outras necessidades (RICHARDSON, 1999). Foram utilizados como dados secundários os documentos e registros do Governo do Estado de Santa Catarina, como: Leis Orçamentárias Anuais dos exercícios de 2012, 2013 e 2014 e do Plano Plurianual 2012-2015.

Baseado nesse conjunto de dados coletados, na legislação em vigor, assim como nas necessidades existentes dos gestores, de controle, avaliação e monitoramento do PPA, surgiu a proposta sistêmica de criação de um portal para o acompanhamento físico e financeiro do Governo do Estado de Santa Catarina, o qual serviu de experiência para a criação da infraestrutura de informação proposta neste trabalho.

3.1 Portal do Acompanhamento Físico e Financeiro do Governo do Estado de Santa Catarina

A transparência estimula a participação social e a informação divulgada aproxima a sociedade da gestão exercida por seus representantes. As entidades públicas têm o dever de promover a transparência de sua administração e a sociedade tem o direito ao acesso e ao acompanhamento da administração pública, como forma de consolidação da cidadania. Segundo Jacobi (2003) para alcançar mudanças na participação social há a necessidade de transformações institucionais que garantam acessibilidade e transparência da gestão.

De acordo com Bobbio (1987), a transparência proporciona um ambiente de análise e reflexão, mas para isso, é necessário que os gestores públicos descortinem suas tomadas de decisões e divulguem-nas livremente nos meios de comunicação acessíveis à população, não permitindo que suas informações fiquem restritas a alguns servidores e assessores.

Portanto, no intuito de fomentar a transparência governamental e o controle social sobre a prestação dos bens e serviços, o Governo do Estado de Santa Catarina desenvolveu durante o exercício de 2013 o Portal do Acompanhamento Físico e Financeiro. Ele possui a finalidade de apresentar de forma simples e objetiva os bens e serviços entregues à sociedade catarinense, utilizando-se de informações do Sistema de Planejamento e Gestão Fiscal (SIGEF). Este sistema concentra todas as informações referentes a execução orçamentária e financeira nos seguintes Módulos: Plano Plurianual, Lei de Diretrizes Orçamentárias, Lei Orçamentária Anual, Execução Orçamentária, Execução Financeira, Contratos, Contabilidade e Acompanhamento Físico e Financeiro.

Além de trazer informações de forma tempestiva e confiável, o Portal possui grande preocupação na linguagem utilizada. Isso porque a transparência pressupõe comunicação eficaz, que por sua vez, pressupõe linguagem adequada e viabilidade de acesso às informações. Nesse sentido, um dos objetivos do portal é o de transformar a linguagem orçamentária e financeira, ou seja, uma linguagem tecnocrata em uma linguagem de fácil compreensão pelo cidadão. É necessário compreender que o controle social não se faz a partir

ISBN 978-85-61115-09-8

da abundância de informações, mas da disponibilidade de informações suficientes e necessárias para o entendimento do cidadão leigo no assunto.

O Portal tem como foco principal o acompanhamento físico e financeiro de todas às ações finalísticas oriundas do Plano Plurianual - PPA vigente que compreende o período de 2012-2015. A Diretoria de Planejamento Orçamentário da Secretaria de Estado da Fazenda considera como ação finalística toda ação geradora de um bem ou serviço prestado a sociedade. Neste sentido, estão em processo de monitoramento 1.411 (hum mil quatrocentos e onze) ações de governo das 3.001 (três mil e uma) ações existentes em todo o plano que financeiramente representam mais de R$ 13,4 bilhões, ou seja, 47% de todas as ações são acompanhadas e as informações disponibilizadas no Portal.

Os objetivos citados nos parágrafos anteriores foram normatizados por meio do Decreto nº 1.324, de 21 de dezembro de 2012, que institui o processo de acompanhamento físico e financeiro e de avaliação do PPA, conforme exposto a seguir:

> Art. 2º O processo de acompanhamento físico e financeiro e de avaliação do PPA tem por finalidade gerar informações que permitam:
> I – divulgar informações de interesse público referentes aos resultados alcançados pela ação governamental;
> II – acompanhar e avaliar os produtos e os resultados alcançados pela ação governamental;
> III – qualificar os processos de elaboração e revisão do PPA, da Lei de Diretrizes Orçamentárias (LDO) e da Lei Orçamentária Anual (LOA);
> IV – corrigir desvios de execução e melhorar a alocação dos recursos públicos; e
> V – subsidiar a elaboração do Relatório de Prestação de Contas do Estado, encaminhado anualmente à Assembleia Legislativa do Estado de Santa Catarina (ALESC) e ao Tribunal de Contas do Estado (TCE).
>
> Parágrafo único. O processo de acompanhamento físico e financeiro e de avaliação do PPA tem como diretriz contribuir para o aprimoramento da gestão pública, da responsabilização, da eficiência, da eficácia e da efetividade dos programas governamentais e do exercício do controle social.

O Portal do Acompanhamento físico e financeiro foi apresentado à sociedade catarinense durante as Audiências Públicas Regionalizadas do exercício de 2014, realizadas pela Assembleia Legislativa do Governo do Estado de Santa Catarina (ALESC), como forma de prestar contas à sociedade dos bens e serviços ofertados a ela nos exercícios anteriores. No entanto, além da prestação de contas, o Portal serviu como uma ferramenta informativa e auxiliadora à sociedade durante o processo de priorização das demandas para o exercício de 2015. Apesar da apresentação do Portal à sociedade seu acesso hoje é possível somente via Intranet do Governo do Estado de Santa Catarina, pois o Portal está em fase de reestruturação para melhor atender as necessidades da sociedade.

3.1 Estudo de Caso: Uso do Portal de Acompanhamento Físico e Financeiro em Audiências Públicas

Desde a implantação do Portal de Acompanhamento Físico Financeiro (PAFF) em Audiências Públicas, em outubro de 2013, a Diretoria de Planejamento Orçamentário da Secretaria da Fazenda orienta a utilização do sistema pelos Órgãos do Poder Executivo, principalmente, como uma ferramenta de gestão e *Business Intelligence* (BI). Assim atende-se um requisito desejado pelos gestores que não encontravam nos relatórios gerados pelo SIGEF as respostas necessárias para a tomada de decisão.

Nesse cenário, foi oportuno a utilização do PAFF durante as Audiências Públicas Regionais realizadas pela Assembleia Legislativa do Estado de Santa Catarina (ALESC) para

ISBN 978-85-61115-09-8

a implantação do conceito de transparência das ações de governo, tanto para a sociedade, quanto para os demais poderes.

No evento realizado em 2014 denominado "Orçamento Regionalizado", os dados referentes ao nível de execução das obras e serviços estabelecidos e previstos no período entre 2012 e 2014 foram apresentados e discutidos entre os participantes. As informações geradas pelos relatórios disponibilizados pelo PAFF permitiram aos participantes:

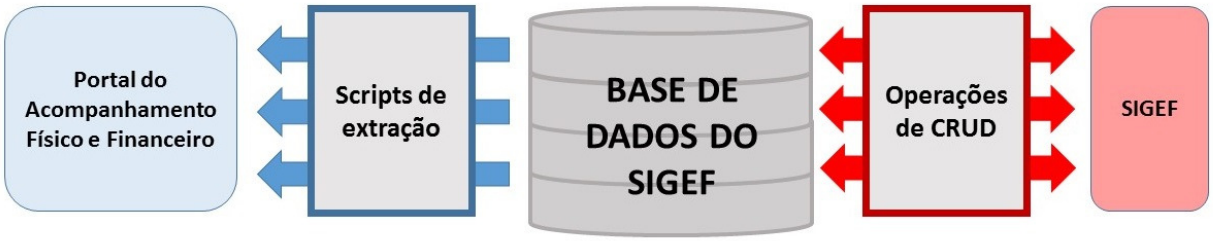
- A análise do percentual de demandas concluídas e atendidas no período;
- A totalização dos valores aplicados em cada Secretaria de Desenvolvimento Regional (SDR);
- A identificação de pendências e impedimentos que geravam os atrasos;
- A distinção entre as atividades em situação de planejamento e execução;
- A visualização por meio de fotografias do real estágio do bem ou do serviço.

Percebe-se que o uso de uma ferramenta para observação das ações de governo é recente e possui muito espaço para evoluir, tanto no aprimoramento dos processos empregados quanto no desenvolvimento de novas funcionalidades. No entanto, a experiência de utilizar o PAFF em um evento de projeção estadual demonstrou que a ferramenta é, de fato, uma grande facilitadora da aproximação da sociedade dos instrumentos básicos que auxiliam no processo de acompanhamento da aplicação dos recursos públicos estaduais. Todavia, devido a grande utilização do Portal pelos diversos órgãos do governo, frequentemente são solicitadas alterações para possibilitar outras visões voltadas à gestão das ações governamentais acompanhadas. Durante este processo de adaptação do Portal ficou evidente que a estrutura inicialmente idealizada não atendia a dinamicidade que a ferramenta exigia, pois pequenas alterações resultaram em um grande esforço da equipe de desenvolvedores para atender as novas exigências. Essa situação, em muitos casos, inviabilizava o atendimento da demanda de forma tempestiva e serviu como o principal estímulo para a criação de uma nova infraestrutura de informação e consequente criação do trabalho

4 Infraestrutura de informação proposta

Conforme mencionado anteriormente, eram evidentes os problemas apresentados pelo Portal do Acompanhamento Físico. Devido à grande quantidade de alterações a que foi submetido, já não apresentava mais boa dinamicidade para atender novas demandas de forma rápida. Além disso, por acessar diretamente a base de dados transacional, oferecia sérios riscos ao desempenho do SIGEF. A figura 2 ilustra como as informações eram obtidas.

Figura 7- ***Forma como os dados eram acessados antes da infraestrutura.***



*Fonte:****Elaborado pelos autores.***

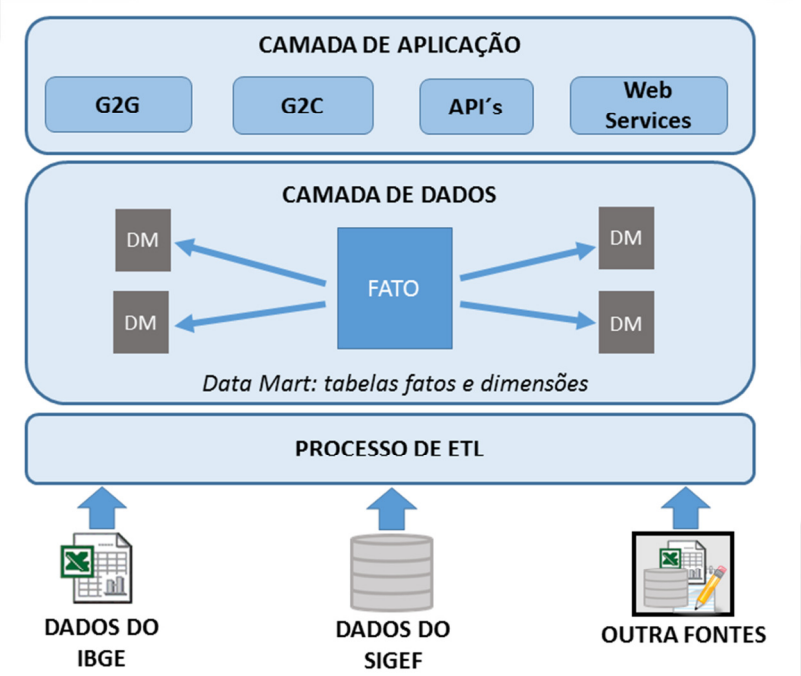
ISBN 978-85-61115-09-8

Como pode ser visto na figura 2, o Portal do Acompanhamento Físico e Financeiro, de acordo com este modelo, busca seus dados acessando diretamente a base de dados do SIGEF. Devido à alta complexidade dos relacionamentos entre as tabelas e da necessidade da sumarização de grandes volumes de dados, aos poucos a implantação de novas funcionalidades foi ficando cada vez mais difícil.

Baseado em tais dificuldades, surgiu a proposta de criação de uma infraestrutura de informação que atenda às necessidades da alta gestão, em tempo hábil, e que possibilite também a disponibilização dos dados conforme às exigências estabelecidas pelo manual de dados abertos governamentais (W3C BRASIL, 2011).

Essa infraestrutura abrange tanto dados do SIGEF quanto de outras fontes de dados. Abrange também processos de ETL (Extraction, Transformation and Load), além de uma camada de dados completamente isolada do SIGEF e uma camada de aplicação. A figura 3 ilustra como a infraestrutura foi construída.

Figura 8 - ***Infraestrutura Proposta***



Fonte: ***Elaborado pelos autores***

A figura 3 ilustra detalhes da infraestrutura proposta, apresentado uma camada de processos de ETL, uma camada de dados e uma camada de aplicação.

Processo de ETL (Extraction, Transformation and Load): consiste no processo de extração de dados das diversas fontes disponíveis, transformação destes para o formato desejado e consequente carga em um data mart presente na camada seguinte. No caso da infraestrutura de informação proposta, os dados são obtidos através (i) do Instituto Brasileiro de Geografia e Estatística (IBGE), através de uma planilha Excel que possui o número de habitantes e o produto interno bruto (PIB) de cada um dos municípios e regioões catarinenses, (ii) do SIGEF e, (iii) de outras fontes, como *scripts SQL* auxiliares ou mesmo qualquer eventual nova fonte de dado que poderá surgir.

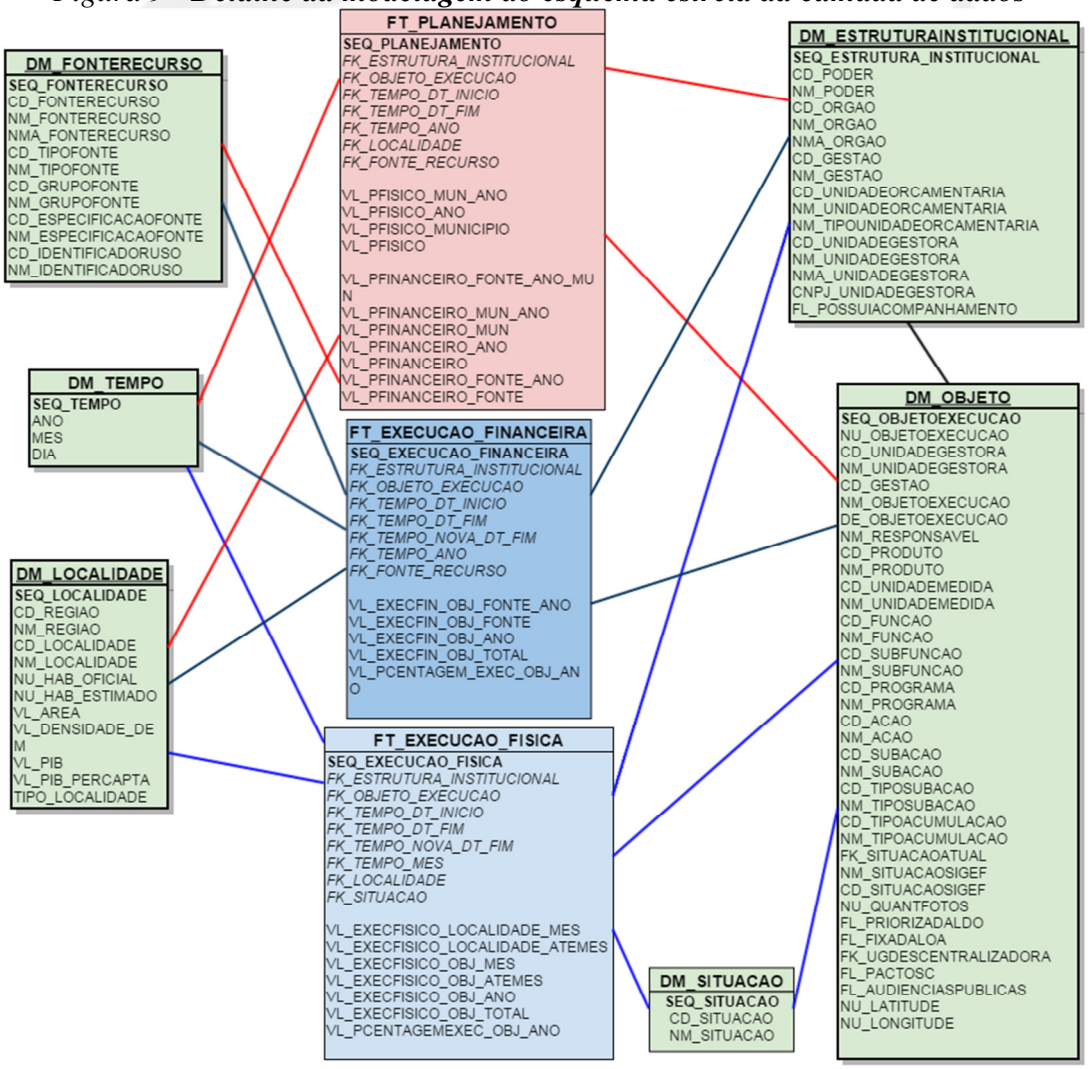
Camada de dados: os dados oriundos do processo de ETL são carregados em um *data mart* formado por tabelas fatos e dimenões. Segundo Kimball (2002), um *data mart* é um conjunto de dados relacionados a um assunto de negócio ou a um departamento; são os dados dispostos em determinada configuração a qual é capaz de representar valores numéricos de

ISBN 978-85-61115-09-8

várias formas distintas estabelecidas pelas dimensões. A figura 4 detalha melhor a camada de dados apresentando a modelagem do *data mart*.

Camada de Aplicação: camada que abrange as aplicações que acessam o data mart proposto. Estas aplicações podem ser criadas tanto para as pessoas (do tipo G2G ou G2C), quanto para máquinas (web services ou mesmo Application Programming Interface - API´s). Estas aplicações também podem ser elaboradas de forma a atender as exigências estabelecidas pelo manual de dados abertos governamentais (W3C BRASIL, 2011), disponibilizando arquivos no formato JSON, CSV, XML, entre outros. Segundo Tauberer (2014), o maior valor dos dados abertos governamentais é proveniente da habilidade da sociedade de realizar suas próprias análises sobre os dados, ao invés de utilizar-se das análises realizadas e publicadas pelo próprio governo. Dessa forma, a camada de abertura de dados consiste na disponibilização de todos os dados na forma bruta para que possam ser utilizados em diversos tipos de aplicações, voltadas para o próprio governo (condição denominada G2G) bem como para os cidadãos (G2C).

Figura 9 - ***Detalhe da modelagem do esquema estrela da camada de dados***



Fonte: ***Elaborado pelos autores***

A figura 4 detalha a camada de dados da infraestrutura de informação, a qual foi exposta na figura 3. Essa camada de dados abrange um data mart com três tabelas fatos e seis tabelas dimensões.

ISBN 978-85-61115-09-8

É detalhada a seguir, uma breve explicação sobre o significado de cada uma das tabelas do *data mart* da figura 4.

- **Dimensão *dm_localidade*:** ou simplesmente dimensão *localidade*, correspondente a umatabela que busca do SIGEF o código e o nome dos municípios e de uma planilha do IBGE o número de habitantes oficial (valor do último censo demográfico) e o seu valor estimado, atualizado anualmente. O PIB e a área também são fornecidos pelo IBGE e os valores per capita são calculados baseados no valor estimado de habitantes.
- **Dimensão *dm_estruturainstitucional*:** é uma tabela que representa toda a estrutura administrativa, desde o poder (executivo, legislativo, judiciário e ministério público), passando pelo órgão até chegar na unidade gestora.
- **Dimensão *dm_fonterecurso*:** para a execução financeira de um objeto, o recurso financeiro pode ser originado de várias fontes diferentes. Hoje existem mais de duzentas fontes cadastradas no SIGEF. Alguns exemplos destas fontes são: recursos do governo federal, recursos do tesouro e fundo social.
- **Dimensão *dm_tempo*:** apenas registros com um identificador juntamente com mais três colunas, dia, mês e ano.
- **Dimensão *dm_situacao*:** possui apenas 7 registros que referem-se à atual situação do objeto de execução:em planejamento, em dia, atrasado, muito atrasado, adiantado, paralisado e concluído.
- **Dimensão *dm_objeto*:** é a tabela de dimensão central do negócio. É nesta tabela que estão todos os atributos como função e subfunção de governo, assim como a definição de se o objeto é projeto ou atividade, se foi descentralizado ou não, se é priorizado na LDO, se é fixado na LOA além de um atributo que define se ele pertence ou não as audiências públicas.
- **Tabelas Fatos:** Em consequência da diferença de granularidade temporal entre a execução e o planejamento de um objeto, decidiu-se separar seus valores em tabelas fatos distintas: *ft_planejamento*, onde os valores registrados na base de dados são de caráter anual, e *ft_execução*, onde os valores - por decisão tomada no processo de levantamento de requisitos - são registrados mensalmente e, de uma forma incremental, para manter assim registros históricos. Mais tarde, por simplificação, viu-se a necessidade de separar essa tabela fato responsável por armazenar os valores da execução do objeto (*ft_execucao*),em outras duas tabelas, *ft_execucao_fisica*, a qual fica responsável por armazenar o histórico da situação do objeto de execução, e a *ft_execucao_financeira*, a qual possui um campo responsável por determinar a fonte de recurso que aponta para a dimensão *dm_fonte_recurso* e é inexistente na tabela.

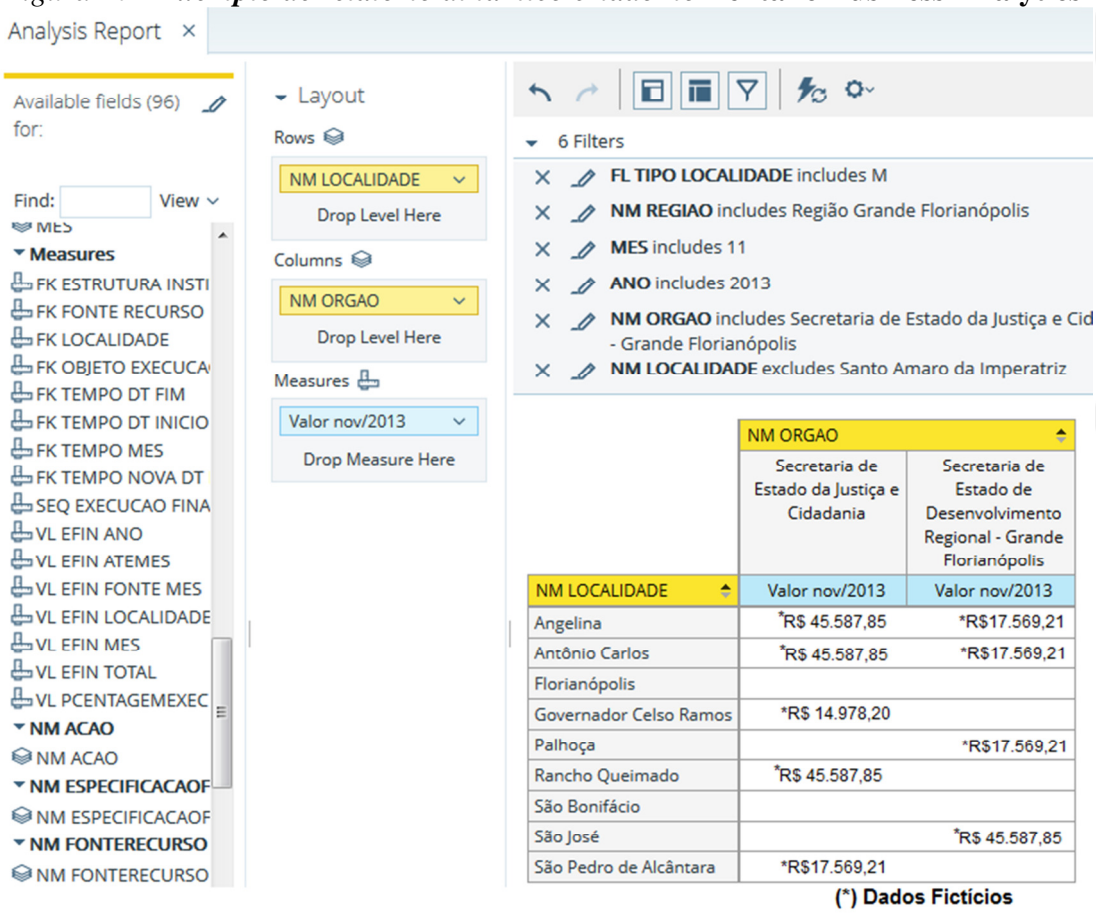
3.1 Nova infraestrutura na prática

Além de problemas de desempenho, a forma anterior de como os dados do acompanhamento físico eram trabalhados não atendia a dinamicidade exigida pelos gestores. A cada nova necessidade de filtros ou de diferentes formatos de visualização, era exigida uma grande demanda de tempo por parte dos desenvolvedores.

Apenas para ilustrar alguns dos problemas praticamente eliminados, foi possível com a nova infraestrutura, a criação de relatórios dinâmicos, analíticos, e até mesmo *dashboards*, dando aos gestores uma liberdade muito maior para elaborarem consultas no momento exato que precisarem. A figura 5 ilustra um exemplo relatório dinâmico criada com a ferramenta *Pentaho Business Analytics*.

ISBN 978-85-61115-09-8

Figura 10 - ***Exemplo de relatório dinâmico criado no*** **Pentaho Business Analytics**



Fonte: ***Elaborado pelos autores.***

A figura 5 ilustra um exemplo de um relatório simples criado com a ferramenta *Pentaho Business Analytics*. Pode-se perceber nela a facilidade com que um usuário pode adicionar ou excluir filtros, adicionar colunas ou mudar métricas.

Considerações Finais

O interesse público e governamental no processo de monitoramento e avaliação dos programas e das políticas públicas está diretamente relacionado à preocupação com a eficácia, a eficiência, a efetividade e a *accountability* de suas ações. O monitoramento e avaliação podem gerar aos gestores públicos informações sobre a qualidade de seu trabalho, como também possibilita demonstrar os resultados à sociedade e aos demais poderes decorrentes do contraste entre o planejado e o realizado.

Nesse sentido, observa-se que a avaliação dos programas governamentais sem um efetivo processo de monitoramento das ações e seus respectivos produtos (bens e serviços), não permite aos gestores públicos intervir no programa durante sua execução com a finalidade de reprogramar as ações para atingir as metas propostas.

Corroborando a esse objetivo, foi criado um portal denominado Portal do Acompanhamento Físico e Financeiro, com o intuito de facilitar o controle, monitoramento e avaliação das políticas públicas por parte da alta gestão governamental. O Portal apresentou grande potencial, não somente para o cumprimento das finalidades para as quais fora criado como também viu-se nele um potencial para sua utilização durante o processo de audiências públicas no Estado de Santa Catarina.

ISBN 978-85-61115-09-8

Durante a adaptação do Portal às novas necessidades impostas pelas audiências públicas, assim como em consequência das novas exigências surgidas à medida que o Portal foi sendo utilizado por mais Órgãos públicos, percebeu-se que as implementações se tornavam cada vez mais insustentáveis e exigiam cada vez mais horas de programação e homologação dos dados.

A ideia de reformular toda a base de dados e criar uma nova infraestrutura de informação surgiu tão logo, não só para melhor atender as necessidades das audiências públicas como também proporcionar aos avaliadores de programas e políticas públicas a possibilidade de relacionar o nível operacional (bens e serviços) com os impactos alcançados pelos programas e políticas públicas. Isso provou-se na prática através da criação de relatórios dinâmicos que vêm sendo utilizado pelos gestores.

Essa nova estrutura foi modelada também para atender os princípios de dados abertos governamentais e possibilitar a transparência e consciência da sociedade frente às ações planejadas e realizadas pelo governo. Uma possível melhora na qualidade das informações e serviços oferecidos pelo governo assim como a colaboração da sociedade de forma mais ativa na gestão pública seriam apenas algumas das consequências dessa infraestrutura.

Como trabalhos futuros, pretende-se realizar as seguintes atividades, a fim de prover informações para análises mais completas e refinadas a respeito das ações governamentais:

- Utilização de outros indicadores do IBGE, relacionados aos municípios e regiões catarinenses;
- Integração com bases de dados de outros sistemas do Estado de Santa Catarina de áreas de atuação como educação, saúde e segurança pública;
- Implementação de uma aplicação para análise e visualização geográfica das informações para disponibilizar aos cidadãos.
- Criação de aplicações como webservices ou API´s que acessam a infraestrutura e disponibilizam os dados para serem consumidos por outras aplicações.

Referências

APPOLINÁRIO, F. **Dicionário de metodologia científica:** um guia para a produção do conhecimento científico. São Paulo: Atlas, 2004.

BARROS, A. J. S. e LEHFELD, N. A. S. **Fundamentos de metodologia**: um guia para a iniciação científica. 2 Ed. São Paulo: Makron Books, 2000.

BRASIL. **Constituição da República Federativa do Brasil**. Brasília, DF: 1988.

BRASIL. **Lei nº 4.320 de 17 de março de 1964** – dou de 23/3/1964. Institui normas gerais de direito financeiro para a elaboração e controle dos orçamentos e balanços da União, dos Estados, dos Municípios e do Distrito Federal. 1964. Disponível em: <http://www.planalto.gov.br/ccivil_ 03/Leis/L4320.htm>. Acesso em: 02 set. 2014.

BOBBIO, Norberto. **Estado governo; por uma teoria geral da política.** 14 ° edição. Rio de janeiro: Paz e Terra, 1987.

CHALFUN, N. **Entendendo a contribuição da política fiscal, do PPA e da LDO para a gestão fiscal responsável, Rio de Janeiro**: IBAM/BNDES 2001.

CONFERÊNCIA WEB W3C BRASIL, 3., 2011, Rio de Janeiro.

GIACOMONI, James. **O orçamento Público.** 15. ed. Revisada e Ampliada, São Paulo: Atlas, 2010.

GIL, Antonio Carlos**. Como elaborar projetos de pesquisa.** São Paulo: Atlas, 1991.

ISBN 978-85-61115-09-8

JACOBI, Pedro. **Educação ambiental, cidadania e sustentabilidade**. Caderno de Pesquisa [online], n.118, p. 189-206, 2003. ISSN 0100-1574.

KIMBALL, Ralph; ROSS, Margy. **The data warehouse toolkit::** the complete guide to dimensional modeling. 2. ed. New York: John Wiley And Sons, 2002.

MATTAR, Fauze Najib. **Pesquisa de marketing**. São Paulo: Atlas, 2005.

SANTA CATARINA. **Constituição do Estado de Santa Catarina**. Florianópolis, SC: 1989.

SANTA CATARINA. **Lei Ordinária nº 15.722, de 22 de dezembro de 2011.** Aprova o Plano Plurianual para o quadriênio 2012-2015 e adota outras providências. São Catarina, Florianópolis: 2011.

SANTA CATARINA. **Decreto Executivo nº 1.324, de 21 de dezembro de 2012.** Institui o processo de acompanhamento físico e financeiro e de avaliação do Plano Plurianual (PPA). Santa Catarina, Florianópolis: 2012.

RICHARDSON, R. J. **Pesquisa social: métodos e técnicas.**3. ed. São Paulo: Atlas, 1999.

TAUBERER, J. **Open Government Data: The Book**. 2 ed. 2014.

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

MÍDIAS DIGITAIS NA FORMAÇÃO DE COMUNIDADES DE PRÁTICA E COMPETÊNCIAS SOCIOEMOCIONAIS

Graziela de Souza Sombrio
graziela.sombrio@ifsc.edu.br

Luiz Antônio Moro Palazzo
luiz.palazzo@gmail.com

Vania Ribas Ulbricht
vrulbricht@gmail.com

Resumo
Este texto tem como objetivo promover uma reflexão sobre o uso de mídias sociais na Educação. Atualmente as mídias fazem parte do cotidiano de milhõesde pessoas e podem ser utilizadas como ferramentas para a melhoria da qualidade de ensino por meio da criação de comunidades de prática e do desenvolvimento de competências socioemocionais. Buscou-se caracterizar o perfil do aluno atual, globalizado e conectado com o mundo com o intuito de criar novas metodologias que utilizem as mídias sociais. Fica evidente que é necessária uma mudança de postura por parte de alunos e professores, pois seus papeis enquanto atores do processo são alterados, podendo ser este o maior desafio encontrado.

Palavras-chave: Mídias sociais. Comunidades de Prática. Competênciassocioemocionais.Educação.

Abstract
This paper aims to promote a reflection about using social media in education. Presently, media take place on the daily life of millions of people and can be used as tools for improving the quality of education through the creation of communities of practice and the development of socio-emotional abilities. We sought to characterize the profile of the contemporary student, globalized and connected with the world aiming create new methodologies using social media. It is observable that a change of attitude coming from the students and the teachers is necessary because their roles while actors of the process are modified, this may be the greatest challenge found.

Key Words: Social media. Communities of Practice.Socio-emotionalabilities. Education.

Introdução

Não é raro no meio docente deparar-se com afirmações como “os alunos não querem nada”, “ os alunos não tem interesse”. Entretanto, essa realidade precisa ser mudada e alternativas precisam ser encontradas.

Em um mundo globalizado, em que o acesso às tecnologias é facilitado, muitas são as opções de diversão na rede mundial de computadores. Redes sociais, filmes online, jogos virtuais, dentre outras tantas possibilidades, fazem parte da rotina da maioria dos jovens de hoje. Será que realmente não há interesse? Ou a forma como as aulas são realizadas não são motivadoras o suficiente para prender a atenção do aluno?

ISBN 978-85-61115-09-8

Segundo Fava (2014, p. 74), os jovens Y[38] e Z[39] querem aprender de forma diferente, pois absorvem informações de forma diferente. Moran (2000) afirma que o aluno "É um cidadão em desenvolvimento. Há uma interação entre as expectativas dos alunos, as expectativas institucionais e sociais e as possibilidades concretas de cada professor".

Fava (2014, p. 79) afirma que "A geração Y encara a globalização como algo normal, até porque suas conexões digitais não reconhecem fronteiras, possuem a liberdade de ir e vir virtualmente a qualquer tempo, em qualquer espaço". Afirma ainda que um estudante Y ou Z sente e leia um livro durante horas pode ser quase inadmissível. (Fava, 2014, p. 74). É um novo perfil, um novo público, uma nova geração. É preciso então buscar alternativas para lidar com as novas gerações, com as novas tecnologias e as possibilidades que surgem constantemente. Isso não significa que tenhamos que abrir mão do rigor, da disciplina, do conteúdo.

1 Mídias na educação

As gerações Y e Z são diferentes. Nem melhores, nem piores, apenas diferentes. A forma de comunicação, de aprendizagem, de leitura e escrita não são mais as mesmas vivenciadas pelas gerações anteriores. Deve-se isso à mudança tecnológica pela qual estamos passando e que deve continuar por um longo período.

Já as escolas, continuam, em grande número, tradicionais, alheias às mudanças que vem ocorrendo. Com isso, um conflito de gerações acontece. Professores utilizando os métodos tradicionais de ensino. Alunos desmotivados e sentindo-se fora do contextto.

A internet, tão utilizada hoje, principalmente por jovens, pode ser utilizada como motivadora para os alunos. Ela passa a ser fonte de pesquisa e ferramenta de interação com seus pares. Moran (2013) nos faz refletir sobre uma das dificuldades encontradas nessa nova forma de ensinar e aprender. Os alunos deixam de ser "recebedores de conteúdo", ou seja, de receber tudo pronto do professor. Com isso, precisam mudar sua "forma de ser" aluno. Já os professores, sentem que não estão "dando aula", pois deixam de repassar o conteúdo tornando assim, as aulas mais interativas.

> Um mundo novo exige uma escola nova e a velocidade das mudanças tem aumentado exponencialmente. Com tantas ferramentas disponíveis para aprender e partilhar, os jovens das novas gerações estão cada vez mais demandando e exigindo das escolas novas posturas e metodologias de ensino. Um modelo de ensino tradicional – ou analógico – não dá conta de suprir as necessidades de um aluno cada vez mais digital. (FAVA 2014, p. XI).

O mundo globalizado em que vivemos força a uma mudança de estratégias e metodologias na educação. Podemos afirmar que os jovens de hoje "são" conectados e têm muito mais acesso à informação do que tínhamos num tempo não tão longíquo.

Fava (2014, p. 51) afirma que "os jovens estudam, trabalham escrevem, aprendem, interagem um com o outro de maneira divergente da sua quando tinha a idade deles". Além das informações, o acesso e a utilização de mídias sociais com a finalidade de "passar o tempo" e se divertir é uma realidade.

[38] Geração Y: nascidos depois de 1983 e antes de 2000.
[39] Geração Z: nascidos depois de 2000.

ISBN 978-85-61115-09-8

Uma definição simples de mídia social, de acordo com Couto, é que se trata da comunicação de todos para todos, em contrário ao que se tinha até bem pouco atrás, onde a comunicação de massa era de um para todos. Com as tecnologias, as mídias digitais, principalmente as sociais, têm feito parte da rotina de milhares de pessoas, especialmente, dos jovens. Eles fazem uso destas ferramentas para suas atividades diárias.

O quadro abaixo mostra exemplos de mídias sociais:

Quadro 1 – Mídias Sociais

Blogs	São páginas na internet voltadas para a disseminação de pensamentos, mas as empresas podem se apoderar desta ferramenta para se relacionarem com o público-alvo que ela almeja.
Redes Sociais	São sites de relacionamento (Facebook, Orkut, Google+, etc) são ótimas ferramentas de divulgação e "viralização", ficam ainda mais poderosas quando são usadas junto com um Blog.
Redes Sociais de Conteúdo	São muito parecidas com as redes sociais normais, apenas focam mais na criação e no compartilhamento de conteúdo, como YouTube, SlideShare, Flickr, etc.
Microblogs	Estas mídias sociais são voltadas para o compartilhamento de conteúdo de forma mais rápida e concisa. (Twitter, Tumblr, Pownce
Mundos Virtuais	Uma Mídia Social ainda pouco explorada, são os simuladores da vida real como o Second Life
Jogos Online	É uma forma muito nova de Mídia Social que ainda não está sendo totalmente explorada. Exemplo: World OfWarcraft

Fonte: COUTO

A figura 1 mostra o que acontece em um minuto na internet, permitindo que se tenha uma noção de como se dá o acesso à rede mundial de computadores, fato que não pode desconsiderar os números apresentados.

ISBN 978-85-61115-09-8

Figura 1 – O que acontece em Um minuto na internet?

Fonte:CAMPOS, 2012

Se as mídias sociais são consideradas motivadoras, por que não utilizá-las para fins educacionais?

A nuvem de termos, da figura 2, mostra elementos que podemos associar ao uso das mídias digitais. Quando se trata de redes sociais, o número é importante – quantas visualizações, quantos amigos, quantos seguidores, quantas "curtidas". Isso mostra o quanto se é "importante" perante à rede. Para o jovem, ser aceito pelos semelhantes sempre foi essencial. Com a rede, tudo se torna mais público.

ISBN 978-85-61115-09-8

Figura 2 – Nuvem de termos

Fonte: MONTENEGRO, 2011

Dentre as formas de acesso às mídias digitais estão os smartphones e outras tecnologias móveis. Desta forma, basta ter acesso à rede para estar conectado, independente do lugar onde estamos.

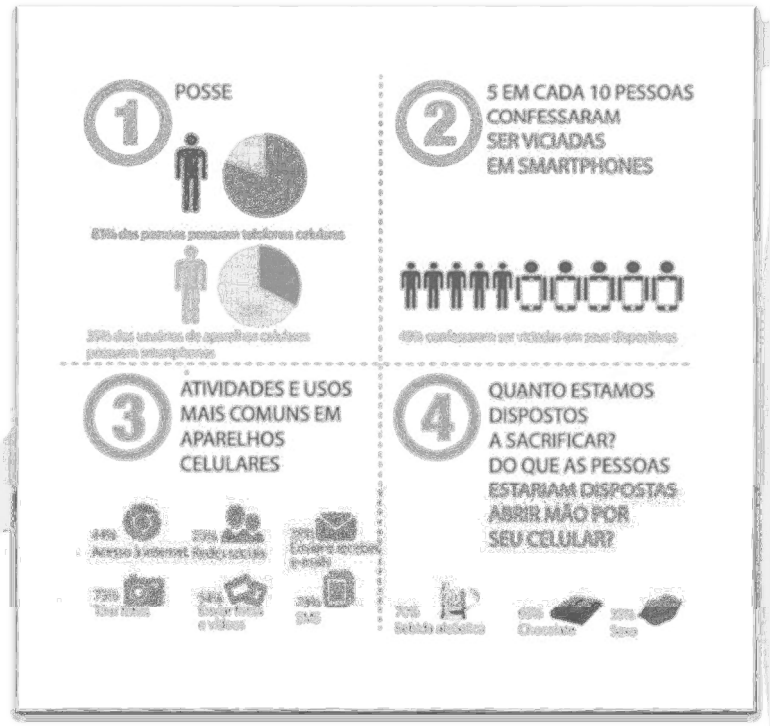
> As tecnologias móveis abrem os horizontes do mundo, as janelas da escola para a vida, a comunicação com múltiplos grupos por afinidade independentemente de onde cada um esteja. (MORAN 2013, p. 58-9)

A figura 3 apresenta dados sobre o uso de telefones celulares, mostrando que o uso de celulares pode, em alguns casos, ser considerado um vício, principalmente entre pessoas jovens. Esse vício pode ser transformado em algo bom e produtivo, se seu uso for direcionado para algo construtivo, como a educação.

.

ISBN 978-85-61115-09-8

Figura 3 – Você é viciado no seu celular?



Fonte:GABRIEL (2013, p. 157).

Terra e Gordon (apud Takimoto, 2014, p. 68) relatam que a união e a participação do indivíduo nas comunidades de prática resultam, em nove benefícios:

1) Aprender com colegas e especialistas;
2) Fazer parte de algo importante e desenvolver uma sensação de identidade;
3) Melhorar o elo com pessoas de outra organização;
4) Desenvolver perspectivas mais amplas da organização e do ambiente;
5) Desenvolver redes pessoais de longo prazo;
6) Receber reconhecimento por habilidades e conhecimentos específicos;
7) Melhorar a autoestima;
8) Novos membros podem facilmente encontrar as principais fontes de conhecimento;
9) Oferecem ambientes para autorrealização e busca de paixões pessoais.

Esses benefícios podem ser alcançados com o uso de mídias sociais na educação como forma de motivação dos alunos, de interação com seus pares e de melhoria no processo de ensino e aprendizagem.

> É fato que a aprendizagem possui alta influência emocional. Também é real que há contágio que diz respeito ao que flui ao longo dos laços de uma rede, pois existe uma irrefutável tendência de os seres humanos influenciarem e copiarem uns aos outros. (FAVA 2014, p. 87)

Muitas intituições de ensino ainda proíbem o uso de celulares e outros dispositivos durante as aulas. Entretanto, essas tecnologias podem servir de aliadas à melhoria da qualidade de ensino. Incorporá-las através de metologias adequadas pode ser uma estratégia para motivar o aluno no processo de ensino e aprendizagem.

ISBN 978-85-61115-09-8

> O ideal é que essas tecnologias *web 2.0* – gratuitas, colaborativas e fáceis – façam parte do projeto pedagógico da instituição para serem incorporadas como parte integrante da proposta de cada série, curso ou área de conhecimento. (MORAN, 2013, p. 33)

Apesar do uso das mídias sociais aplicadas à educação ser um tema relativamente novo, uma pesquisa no Portal da Capes (tabela1) revela os seguintes dados:

Tabela 1 – Quantitativo de trabalhos encontrados no Portal da Capes

Termos pesquisados	Quantitativo de artigos encontrados	Artigos revisados por pares	Quantitativo de artigos considerando período de publicação	Realizado filtro pela tema do periódico (sim ou não)	Quantitativo de artigos por idioma	Selecionados pelo título, considerando a relevância para o presente estudo
Mídias sociais + educação	209	90	Publicados de 2010 em diante: 51	Não	Inglês – 4 Espanhol – 46	9 artigos
Blog + educação	152	54	Publicados de 2010 em diante: 41	Não	Inglês – 13 Espanhol – 27 Português - 1	10 artigos
Redes sociais + educação	1219	646	Publicados após 2011 em diante: 189	Sim – restaram 13 artigos	Espanhol – 11 Inglês - 2	4 artigos
Mundo Virtuais + educação	399	219	Publicados após 2009: 130	Sim – restaram 21 artigos	Espanhol – 17 Inglês - 4	6 artigos
Jogos on line + educação	64	26	Publicados após 2010: 12	Não	Espanhol – 10 Inglês - 2	2 artigos

Fonte:produzido pelos autores

Trabalhos relativos às midias sociais aplicadas à educação têm sido realizados, o que mostra a importância do tema, principalmente nos dias atuais. Pelo quadro acima, percebe-se que em torno de 50% das publicações são recentes e que a grande maioria dos trabalhos são de língua espanhola. São novas possibilidades de integração que buscam o mesmo objetivo: a melhoria da Educação.

As mídias sociais podem ser utilizadas como espaços colaborativos para a construção do conhecimento. Por exemplo, alunos de uma mesma turma podem trocar ideias sobre exercícios, postar materiais de apoio e trocar experiências. Com isso, grupos de estudos virtuais são criados para colaborar com o processo de ensino e aprendizagem. O papel do professor é de mediador (moderador), acompanhando as atividades realizadas. Isso é essencial para que o processo seja feito com seriedade e sem perda dos objetivos pedagógicos.

> A finalidade das atividades do moderador é promover a criação de inteligência coletiva em rede e manter a coesão, o que na prática significa qualidade de interação, o nível necessário de cooperação e construção do conhecimento alcançado (GAIRÍN-SALLAN; RODRIGURS-GOMES; ARMENGOL-ASPARÓ, 2010 apud TAKIMOTO, 2014).

Quando nos reportamos às mídias sociais como recurso metodológico para o processo de ensino e aprendizagem, precisamos relacionar ao conceito de Comunidade de Prática.

ISBN 978-85-61115-09-8

Nestas comunidades, pessoas com um interesse ou tema em comum se unem com o propósito de resolução de um problema.

> O conceito de Comunidade de Prática foi construído justamente em torno da atividade, onde um grupo de indivíduos com interesses comuns em um dado domínio,compartilham práticas mutuamente negociadas, crenças, compreensões, opiniões, valores e comportamentos. (VANZIN, 2005, p. 37)

No universo das Comunidades de Prática os processos de aprendizagem podem ser cooperativos e colaborativos. Para Vanzin (2005), apesar destes conceitos se confundirem, eles se completam quando tratamos de Comunidades de Prática. Para o autor,

> A diferença está na forma de realização da atividade, sendo a cooperação realizada pela divisão do trabalho entre participantes como uma atividade em que cada pessoa é responsável por uma porção da resolução do problema. A colaboração tem sua identidade no engajamento mútuo dos participantes em um esforço coordenado para juntos resolverem o problema. (VANZIN, 2005, p. 40).

As mídias digitais podem ser utilizadas como Comunidades de Prática em um processo de aprendizagem colaborativa. Atráves das mídias, as informações podem ser compartilhadas, acessadas de qualquer lugar, a qualquer momento. O professor, além de mediador, passa a ser também um membro colaborador. A construção é função de todos, de forma comprometida com o próposito comum. Pode ser vista como uma união de competências individuais para gerar resultados e valores coletivos.

> A educação escolar precisa compreender e incorporar mais novas linguagens, desvendar os seus códigos, dominar as possibilidades de expressão e as possíveis manipulações. É importante educar para usos democráticos, mais progressistas e participativos das tecnologias, que facilitem a evolução dos indivíduos. (MORAN, 2013, p. 53)

No mesmo mundo globalizado, que apresenta novas tecnologias, novas possiblidades de comunicação, o que se espera dos profissionais em todas as áreas do conhecimento não são mais apenas as habilidades cognitivas. Espera-se pessoas que saibam utilizar esses conteúdos, que sejam líderes, criativos, colaboradores, empreendedores, entre tantas outras características. Para isso, é preciso desenvover as capacidades não-cognitivas, chamadas também de habilidades socioemocionais.

O Instituto Ayrton Senna em parceria com a OCDE, agrupou seis habilidades socioemocionais mais importantes para serem avaliadas (CHAN, 2014). São elas:

1) Determinação – Para ter determinação é preciso organização, disciplina e foco.
2) Colaboração – Dificilmente uma pessoa trabalha sozinha. A colaboração é uma das características que se espera de um profissional.
3) Sociabilidade – É a capacidade de interagir com outras pessoas, tornando o ambiente agradável e produtivo.
4) Estabilidade emocional – É saber lidar com ambientes e situações de estresse e de dificuldade.
5) Protagonismo – Ser protagonista da própria vida.
6) Curiosidade – A curiosidade leva à criatividade e à inovação.

ISBN 978-85-61115-09-8

Todas essas características podem e devem ser levadas ao ambiente escolar. Para termos adultos com capacidades socioemocionais bem desenvolvidas pode-se começar desde a escola, sem limite mínimo de idade. Além disto, as competências e habilidades socioemocionais estão em consonância com as caracteristicas desenvolvidas por meio das comunidades de prática.

As competências socioemocionais também podem ser devenvolvidas por meio das mídias digitais. A melhoria da educação passa pelo desenvolvimento dessas competências, formando pessoas críticas, criativas e inovadoras.

Considerações Finais

O Plano Nacional de Educação prevê como uma estratégia para a melhoria da aprendizagem dos alunos, o desenvolvimento de tecnologias educacionais e práticas pedagógicas inovadoras. Para isso, o documento traz como meta (16), o aperfeiçoamento permanente dos professores da educação básica em sua área de atuação, considerando os avanços no campo educacional. Por outro lado, as Diretrizes Curriculares Nacionais para a Educação Básica prevêem a utilização de novas tecnologias educacionais, como processo de dinamização dos ambientes de aprendizagem. Reforçam que é necessária a formação adequada aos professores e que o número de recursos midiáticos esteja de acordo com o número de alunos.

Sem sombra de dúvida, as mídias sociais são uma alternativa para a utilização de tecnologias na educação. Além de uma mudança na função do professor e na postura do aluno, é preciso um investimento por parte das instituições de ensino. Acreditar que é possível é o primeiro passo. Ambientes adequados, estrutura de computadores e de internet são essenciais para a implantação dessas novas tecnologias. Fazer essa implantação é uma forma de criar comunidades de prática com objetivos educacionais, possibilitando também o desenvolvimento de competências socioemocionais.

Referências

Brasil. **Diretrizes Curriculares Nacionais Gerais da EducaçãoBásica** / Ministério da Educação. Secretaria de EducaçãoBásica. Diretoria de Currículos e Educação Integral.Brasília: MEC, SEB, DICEI, 2013.562p.

BRASIL. Lei nº 13005, de 25 de junho de 2014. **Plano Nacional de Educação**. Disponível em: <http://www.planalto.gov.br/CCIVIL_03/_Ato2011-2014/2014/Lei/L13005.htm>. Acesso em: 14 set. 2014.

CAMPOS, George. **Infográfico - O que acontece em um minuto na internet?** 2012. Disponível em: <http://www.georgecampos.me/blog/tecnologia-da-informacao/infografico-o-que-acontece-em-um-minuto-na-internet/>. Acesso em: 12 set. 2014.

CHAN, Iana. **O que são competências não-cognitivas?** 2014. Disponível em: <http://educarparacrescer.abril.com.br/aprendizagem/sao-competencias-nao-cognitivas-777484.shtml>. Acesso em: 13 set. 2014.

COUTO, Guilherme. **O Que é Mídia Social?** Disponível em: <http://www.marketingdigitaldicas.com.br/o-que-e-midia-social>. Acesso em: 12 set. 2014.

ISBN 978-85-61115-09-8

DUBNER, Deborah. **Mídias sociais: você está preparado?** Disponível em: <http://www.midiasocial.com.br/home/midias-sociais.asp>. Acesso em: 12 set. 2014.

FAVA, Rui. **Educação 3.0:** aplicando o PCDA nas instituições de ensino. São Paulo: Saraiva, 2014.

GABRIEL, Martha. **Educ@r:** a (r)evolução digital na educação. São Paulo: Saraiva, 2013.

MONTENEGRO, Chico. **Engajamento, Sentimento e Influência são as métricas mais usadas por analistas de mídias sociais e monitoramento.** 2011. Disponível em: <http://midiaboom.com.br/dados-e-estatisticas/engajamento-sentimento-e-influencia-sao-as-metricas-mais-usadas-por-analistas-de-midias-sociais-e-monitoramento/>. Acesso em: 12 set. 2014.

MORAN, José. **Mudar a forma de ensinar e de aprender.** 2000. Disponível em: <http://www.eca.usp.br/prof/moran/site/textos/tecnologias_eduacacao/uber.pdf>. Acesso em: 13 set. 2014.

MORAN, José Manuela; MASETTO, Marcos T.; BERHENS, Marilda Aparecida. **Novas tecnologias e mediação pedagógica.** 21. ed. Campinas: Papirus, 2013. 171 p. (Coleção Papirus Educação).

TAKIMOTO, T. **A percepção do espaço tridimensional e sua representação bidimensional:** a geometria ao alcance das pessoas cegas em comunidades virtuais de aprendizagem. Florianópolis: UFSC/Engenharia e gestão do conhecimento, 2014. Dissertação de mestrado.

VANZIN, T. **TEHCo – Modelo de ambientes hipermídia com tratamento de erros, apoiado na teoria da cognição situada.** Florianópolis: UFSC/Engenharia de Produção, 2005. Tese de doutorado.

ISBN 978-85-61115-09-8

O VALOR DOS DADOS ABERTOS LIGADOS: PROPOSTA DE AVALIAÇÃO

Silvia Maria Puentes Bentancourt
silviampb@gmail.com

Denise Santin Ebone
deniseebone@gmail.com

Dr. Rogério Cid Bastos
rogerio@egc.ufsc.br

Resumo
Discute o valor que dados abertos disponíveis na *web* ganham ao estarem ligados. Aborda a ligação entre dados desde a perspectiva de um serviço, onde a avaliação inclui ativos intangíveis. Utiliza para a discussão os dados oferecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE) pelo volume e valor dos mesmos. Apresenta as contribuições dos dados abertos e ligados em relação àqueles que são somente abertos, baseado em uma pesquisa exploratória, qualitativa e bibliográfica.

Palavras-chave: Dados abertos (*Open Data - OD*). Dados ligados (*Linked Data - LD*). Dados abertos ligados (*Linked Open Data - LOD*). IBGE. Oportunidade de uso.

Abstract
This paper discusses the value earning of the available open data on the web when connected. Addresses the link between data from the perspective of a service, where the assessment includes intangible assets. It use for discussing the data offered by the Instituto Brasileiro de Geografia e Estatistica (IBGE) due to its volume and value. Presents the contributions of linked open data towards those that are only open, based on an exploratory, qualitative and literature research.

Keywords*: Open Data (OD). Linked Data (LD). Linked Open Data (LOD).IBGE.Use Opportunity.*

1 INTRODUÇÃO

O conhecimento cresce quando compartilhado e as organizações têm capacidade de aprender umas com as outras de acordo com estratégias, intenções e procedimentos definidos. A aprendizagem organizacional é um processo dinâmico de criação de conhecimento e da sua transferência para onde é necessário (KANE; ALAVI, 2007). As tecnologias disponíveis oferecem possibilidades para extração, mineração e aproveitamento do conhecimento, de forma colaborativa na *web*. Sendo o conhecimento colaborativo fundamental para criação, intercâmbio e transformação do conhecimento (JONES, 2001).

O conceito de dados abertos (*Open Data*) é entendido a partir da proveniência e da necessidade do usuário para responder às suas questões. A proveniência possibilita a avaliação da confiabilidade dos dados (ALI *et al.*, 2013). Quanto às demandas do usuário, o fato de dispor dados abertos na *web* permite que sejam tratados de forma a propiciar melhor

ISBN 978-85-61115-09-8

uso. Assim, eles são codificados novamente, reorganizados e analisados a fim de trazer novas contribuições e aplicações, elevando-se o seu valor ao patamar do conhecimento.Com melhorias na tecnologia, ferramentas e comunicações, os dados tornaram-se muito mais fáceis de coletar, armazenar, gerir, distribuir e reutilizar (WALLIS *et al.*, 2013). Trata-se da aplicação prática da espiral do conhecimento preconizada por Nonaka e Takeuchi (1997), de forma aberta. Então, os dados são disponibilizados, podendo oferecer informação e novo conhecimento de modo contextualizado, resultando em transformações e benefícios à sociedade pela agregação de valor ao conhecimento. A ampliação acontece, ainda, pelas variadas possibilidades de aplicação dos dados abertos, de acordo com os agentes humanos em função de seus interesses, associações, contextos e conhecimentos prévios.

Portanto, é interessante queos dados sejam descritos de forma a especificar exatamente o que eles representam, que sejam evidenciados os critérios de coleta, validação, apresentação, dentre outros. Algumas dessas informações poderão estar presentes nos metadados, outras de forma explícita e textual. Apenas disponibilizar os dados não é suficiente, fazendo-se necessário o tratamento e a sua correta descrição, que ainda poderá auxiliar na sua localização de forma rápida e eficiente.

A *Web* Semântica favorece inicialmente a estruturação, ressignificação e integração dos dados disponibilizados, de forma automática, por meio de ferramentas tecnológicas.Ferraram *et al.* (2013) estudaram uma série de problemas relativos à *Web* Semântica, sendo os principais a comparação e combinação de dados e a capacidade de resolver a multiplicidade de referências de dados para os mesmos objetos do mundo real. Portanto o passo seguinte foi o desenvolvimento de ambientes interativos para correlacionar os diferentes conjuntos de dados, facilitando as novas oportunidades de uso, novas aplicações e manipulação por diferentes ferramentas. As correspondências de dados ocorrem na forma de *links* de dados (FERRARAM *et al.*, 2013), os dados ligados (*Linked Data*) referem-se aos dados publicados na *web*, de forma legível para uma máquina, com seu significado explicitamente definido eligado a outros conjuntos de dados externos (BIZER *et al.*, 2009).

A junção dos conceitos anteriores recebeu a denominação *LinkedOpenData* (LOD) termo original no idioma inglês. Desta forma, passou-se da *Web* de Documentos para a *Web* de Dados. No primeiro caso, os documentos, que são dados não estruturados, se relacionam na *web* pelos *hiperlinks*, onde o usuário escolhe o percurso adequado a suprir suas necessidades informacionais. Enquanto que na *Web de Dados*, os agentes não humanos ligam os dados de forma a permitir seu entendimento e aplicação (BIZER *et al.*, 2008).

Por meio dos dados abertos ligados, um mesmo conjunto de dados pode ser relacionado a outro de forma particular, em função de um propósito. Os novos conhecimentos que auxiliam as tomadas de decisão oriundos dessas novas relações podem levar à confirmação de saberes conhecidos, mostrar novos usos a partir de fatos conhecidos e, até mesmo, evidenciar lacunas de conhecimento.

A ligação dos dados pode ser vista como um serviço, considerando-se elementos intangíveis como acessibilidade e confiabilidade, onde se agrega valor aos dados já abertos. Cabe salientar que, se por um lado existem benefícios, a não disponibilização de dados ligados na sociedade do conhecimento pode acarretar efetivas perdas de oportunidades e de vantagens competitivas, o que é passível de mensuração com as ferramentas adequadas.Assim, busca-se identificar critérios que norteiem a avaliação do valor agregado aos dados abertos ligados.

O presente artigo está dividido em 5 sessões, sendo a primeira a introdução, seguido de um referencial teórico que aborda os dados abertos ligados, dados abertos governamentais, o valor agregado aos dados abertos quando ligados e a ligação entre dados abertos entendida como um serviço. Segue um breve histórico do IBGE, que foi escolhido para servir de estudo

ISBN 978-85-61115-09-8

de caso, e se discute uma proposta de avaliação dos dados abertos ligados. Finaliza com algumas considerações derivadas do estudo.

2 REFERENCIAL TEÓRICO

A fim de discutir o valor dos dados abertos ligados, apresentam-se alguns conceitos que servirão de apoio nesse sentido.

2.1 Dados Abertos Ligados (LOD)

Dados abertos são aqueles que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer um (OPEN DEFINITION, *online*). Existem características importantes que os dados devem ter para serem considerados abertos (OPEN DEFINITION,*online*):

1. acesso: os dados devem ser disponibilizados na íntegra, de preferência com *download* através da internet sem representar custo;
2. redistribuição: a licença não deve restringir ninguém de vender ou disponibilizar os dados sem custo algum;
3. reutilização: a licença deve permitir modificações e trabalhos derivados, também deve ser permitida a distribuição sob os mesmos termos do trabalho original;
4. ausência de restrições tecnológicas: os dados devem ser apresentados de forma que não existam obstáculos tecnológicos;
5. atribuição: a licença pode exigir como condição para a redistribuição e reutilização a atribuição dos colaboradores e dos criadores da obra;
6. integridade: a licença pode exigir, como condição para a obra que está sendo distribuída de forma modificada, a identificação de um nome ou número de versão diferentes do trabalho original;
7. sem discriminação contra pessoas ou grupos: a licença não deve discriminar qualquer pessoa ou grupo de pessoas;
8. sem discriminação contra campos de trabalho: a licença não deve restringir ninguém de fazer uso do trabalho em um campo específico de atuação;
9. distribuição da licença: os direitos associados ao trabalho devem ser aplicados para todos aqueles a quem o trabalho é redistribuído.
10. licença não deve ser específico para uma coletânea: os direitos associados à obra não devem depender do trabalho ser parte de uma coletânea específica; e
11. licença não deve restringir a distribuição de outras obras: a licença não deve colocar restrições em outros trabalhos que são distribuídos juntamente com as obras licenciadas.

Essa série de requisitos tem como finalidade orientar o uso, reuso e distribuição dos dados. O termo licença refere-se às condições legais sob as quais o trabalho é disponibilizado (OPEN DEFINITION, *online*). Nos dados abertos, a única restrição permitida é a identificação da fonte e que o compartilhamento ocorra seguindo as mesmas regras (OPEN KNOWLEDGE FOUNDATION, 2012).

Existem também os dados ligados, *linked data*foi a denominação proposta por Tim Berners-Lee (2006) para designar uma forma de publicação de dados que permite o intercambio de dados estruturados com a mesma facilidade com que se trocam documentos ou direciona-se de uma página a outra na *web*. *Linked data* trata da junção de dados guardados em bases de dados diferentes, onde se direciona de uma base à outra sem necessariamente ter

ISBN 978-85-61115-09-8

controle sobre nenhuma delas. Adota-se a premissa na qual o valor e a capacidade de uso de um dado crescem quanto mais estiver conectado a outro dado (BIZER; HEATH; BERNERS-LEE, 2009).

Nessa lógica, a *web* pode chegar a ser uma grande base de dados, com uma infindável possibilidade de cruzamentos. A ligação dos dados constitui-se na estrutura para que estes sejam descobertos, acessados, integrados, utilizados e reutilizados com facilidade. São quatro os princípios que regem os dados ligados:

1. uso de um identificador único (*Universal ResourceIdentifiers* - URI) como nome de objetos, sejam coisas ou conceitos;
2. uso de HTTP URIs para acessar àqueles nomes;
3. quando acessado um URI, que se encontrem mais informações úteis e que sejam utilizados padrões para isso; e
4. inclusão de links para outros URIs que facilitem encontrar mais coisas.

Berners-Lee (2009) chama a atenção que não se trata de regras que não possam ser quebradas, mas de recomendações a serem seguidas para não perder a oportunidade de ter dados interconectados. Logo, éum conjunto de boas práticas de publicação e conexão de dados na *web*, onde se busca evitar a ambiguidade ao usar o URI. As conexões podem ocorrer pelo objeto que os dados representam ou por suas propriedades e conceitos (BERNERS-LEE; O'HARA, 2013). Desse modo,por exemplo, dados sobre uma universidade poderão ser relacionados a outras instituições de ensino, cursos, currículos, assim como à cidade onde está localizada, indicadores econômicos, demográficos e assim por diante. Um único conjunto de dados poderá ter ilimitadas ligações com outros conjuntos.

Porém, para ligar os dados distribuídos na *web* é necessário um mecanismo padrão para especificar o significado das conexões entre os itens descritos nos dados.Diante disso, procurou-se estabelecer padrões.O padrão utilizado é o RDF (*ResourceDescription Framework*), que permite apresentar os dados de forma que as máquinas consigam compreender e interpretar as informações adicionais sobre eles. Este tipo de prática caracteriza a *Web de Dados*, em que o propósito principal é, através de tecnologias de informação e comunicação, extrair conhecimento a partir de dados disponíveis na *web*. Portanto, o usuário pode partir de uma fonte de dados e explorar através de conexões com padrão RDF a infindável *web* de dados (BIZER *et al.*, 2008).

2.2 Dados Abertos Governamentais

Dados abertos constituem-se num ótimo recurso para adquirir conhecimento. Entre as áreas atualmente favorecidas pelos dados abertos temos (GRAY *et al.*, 2011):

- transparência e controle democrático;
- participação popular;
- empoderamento dos cidadões;
- melhores ou novos produtos e serviços privados;
- inovação;
- melhora na eficiência de serviços governamentais;
- melhora na efetividade de serviços governamentais;
- medição do impacto das políticas; e
- conhecimento novo a partir da combinação de fontes de dados padrões.

ISBN 978-85-61115-09-8

Estas dimensões potencializam o uso de dados governamentais quando abertos. Entende-se por dados abertos governamentais aqueles produzidos ou disponibilizados por instituições governamentais ou controladas por ele e que seguem os princípios de dados abertos (OPEN KNOWLEDGE FOUNDATION, 2012). Tais dados são públicos e podem serentendidos como de propriedade do cidadão (DINIZ, 2010).

Com o propósito de disponibilizar dados governamentais, a política nacional de dados abertos está configurada na Infraestrutura Nacional de Dados Abertos (INDA). A INDA "é um conjunto de padrões, tecnologias, procedimentos e mecanismos de controle necessários para atender às condições de disseminação e compartilhamento de dados e informações públicas no modelo de Dados Abertos, em conformidade com o disposto na e-PING"(PORTAL BRASILEIRO DE DADOS ABERTOS, *online*). A e-PING constitui-se na arquitetura de interoperabilidade do governo brasileiro.

Também como parte da INDA, foi criado o Portal Brasileiro de Dados Abertos (dados.gov.br)com o propósito de permitir a reutilização dos dados governamentais e propiciar maior interação com a sociedade. Ele é mais uma ação no sentido de facilitar o acesso à informação, como prescrito na Lei n. 12.527, de 18 de novembro de 2011, que trata propriamente do acesso à informação.

A princípio, no portal faz-se a convergência de todos os dados publicados pelo governo federal, mas já explicita a intenção de ampliar para as esferas estaduais e municipais. Os mecanismos de busca do portal permitem encontrar os dados pela fonte de publicação, área a que se referem, tipo de licença de uso, tipo de arquivo, assim como por termos específicos.

2.3 Valor dos Dados Abertos Ligados (LOD)

Dados abertos também têm grande importância econômica, que pode aumentar se os dados forem ligados. Estudos estimaram que seu valor monetário seja de dezenas de bilhões de euros, somente na União Europeia, onde novos produtos de novas empresas estão reutilizando esses dados e gerando valor com eles (GRAY*et al.*, 2011).

Os governos armazenam dados em muitas áreas que são de interesse direto para os cidadãos, empresas e outros consumidores de dados, incluindo outras agências governamentais. Exemplos incluem o planejamento da cidade, tráfego, dados administrativos, meio ambiente, educação, informação lazer, infraestrutura e muitos mais (KASCHESKY; SELMI, 2014). Esses dados abertos tem o potencial para ser reutilizado levando a criação de novos produtos e serviços. A seguir, alguns exemplos de projetos realizados utilizando dados abertos ligados:

- transparência - projetos como o *taxtree* finlandês e o *where does mymoney go* britânico mostram como o dinheiro dos impostos está sendo gasto pelo governo; Aplicativos que mostra as atividade do parlamento como *zewerkenvoorjou.be* e *TheyWorkForYou.com;*
- qualidade de vida - serviços como o *mapumental* no Reino Unido e o *mapnificent* na Alemanha permitem que encontrar locais para morar, considerando a duração de translado ao trabalho e também verifica os preços das moradias; e
- meio ambiente - o *husetsweb.dk* dinamarquês ajuda a encontrar alternativas para melhorar a eficiência energética da casa, incluindo o planejamento financeiro, e apresenta construtores que podem fazer o trabalho (OPEN KNOWLEDGE FOUNDATION, 2012).

ISBN 978-85-61115-09-8

Assim, novas combinações de dados também podem criar novos conhecimentos e descobertas, que conduzem a campos de aplicação totalmente novos. No âmbito de decisões governamentais, um exemplo, foi a descoberta da relação entre a poluição da água potável e a cólera, feita pelo Dr. Snow em Londres no século XIX. Ele combinou dados sobre mortes devido à cólera com a localização das cisternas d'água. Seus resultados contribuíram na construção do sistema de esgoto de Londres e, por conseguinte, na melhoria da saúde geral da população (OPEN KNOWLEDGE FOUNDATION, 2012).

Em vista disso, novas descobertas como essa podem acontecer, na medida em que revelações inesperadas surgirem da combinação de diferentes conjuntos de dados.Portanto o valor dos dados abertos ligados passa a ser exponencial na medida em que mais dados forem disponibilizados dessa forma na *web*.

2.4 Dados Abertos Ligados (LOD) como Serviço

O conhecimento é alcançado através da organização e uso da informação que se originam em dados diversos. Tal diversidade surge pelos diferentes critérios que podem ser classificados, como a fonte que os disponibilizam, o grau de agregação, o tipo de informação que representam, entre outros. Assim, uma maneira de classificar é quanto à liberdade de uso e reuso dos dados, o que resulta em dados fechados, parcialmente fechados e abertos. Os fechados e parcialmente fechados permitem acesso a usuários com algum tipo de autorização e não fazem parte do escopo deste estudo. Já os dados abertos, ganham interesse em função de suas aplicações, principalmente quando estes estão ligados a outros incrementando seu valor.

Nesse contexto, assim como há os produtores e distribuidores de dados abertos ligados, existem os consumidores. Para que tudo possa ocorrer com qualidade, deverão existir instrumentos que permitam a correta transmissão, leitura e apresentação dos dados a fim de extrair informação deles. Trabalhos nesse sentido tem-se desenvolvido junto a *Web Semântica* e iniciativas como a *DBpedia*, que depois foi seguida por instituições públicas e privadas, inclusive de informações estatísticas (BERNERS-LEE; O'HARA, 2013).

Segundo Diniz (2010) não há valor na disponibilização de dados governamentaisabertos se a sociedade não tem interesse em reutilizá-los. Mas tal afirmação também alerta para os cuidados que devem ser tomados na forma de apresentar os dados, pois um dos motivos da falta de interesse pode ocorrer pela dificuldade de acesso à informação que se deseja.

O exposto remete a que propiciar facilidade de uso é um meio de agregar valor àqueles dados. Avaliar decisões comparando custo e benefício são comuns nas organizações e pode servir também para conhecer o valor dos dados abertos ligados, mais precisamente, por métodos de avaliação da qualidade de serviços.

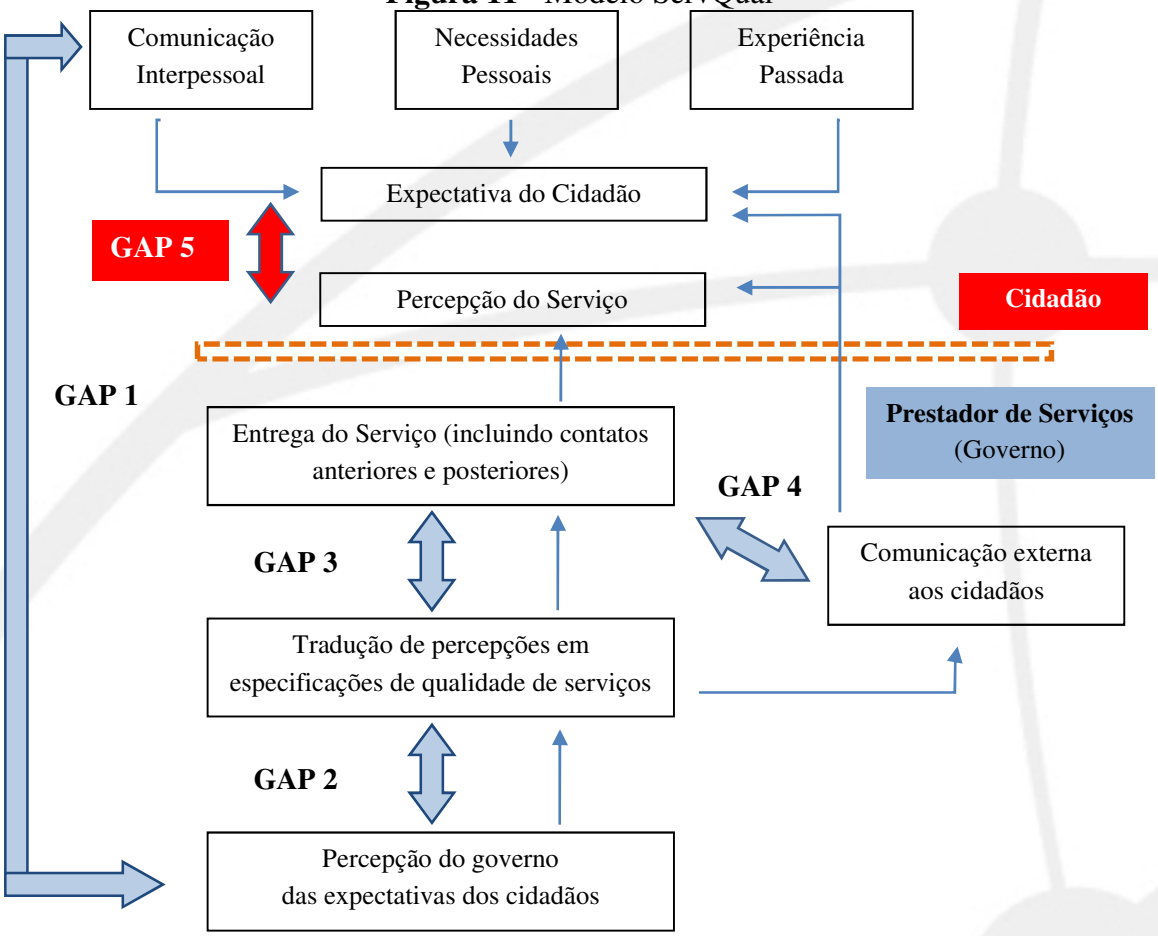
Segundo Albrecht (2000, p. 50) "Serviço é todo trabalho feito por uma pessoa em benefício de outra. E por qualidade compreende-se que é a medida que uma coisa ou experiência satisfaz uma necessidade, soluciona um problema ou agrega valor em benefício de uma pessoa". A ligação entre dados abertos pode ser considerada como um serviço e, sendo assim, uma maneira de estimar o seu valor seria através dos elementos que caracterizam um serviço, mais precisamente, daqueles que avaliam a qualidade de um serviço. Estes incorporam aspectos tangíveis e intangíveis, onde a avaliação da qualidade é extremamente subjetiva, baseada na expectativa e na percepção de quem recebe o serviço.

Um dos métodos mais utilizados para avaliar a qualidade de um serviço é o proposto por Parasuraman, Zeithaml e Berry (1985), denominado ServQual. Nele se evidenciam 5 lacunas (*gaps*) que são as diferenças entre o esperado e o percebido pelo usuário ou cliente, sendo que também envolve as expectativas e percepções dos gestores e prestadores do serviço. As expectativas podem ser não atendidas, atendidas ou excedidas. O mesmo modelo

ISBN 978-85-61115-09-8

é utilizado para avaliar a qualidade da informação. Neste estudo, foi adaptado alterando o cliente por cidadão e os gestores e prestadores de serviço pela figura do governo. As 5 lacunas e sua interação são mostradas na Figura 1.

Figura 11 - Modelo ServQual



Fonte: adaptado de Parasuraman, Zeithaml e Berry (1985)

Cada lacuna corresponde a um tipo de discrepância, qual seja:

- lacuna(*gap*) 1: discrepância entre as expectativas dos cidadão e as percepções dos prestadores de serviços, agentes do governo, sobre essas expectativas;
- lacuna(*gap*) 2: discrepância entre a percepção dos prestadores de serviços, agentes do governo, em relação às expectativas dos cidadãos e a especificação do serviço;
- lacuna(*gap*) 3: discrepância entre a especificação do serviço e os serviços realmente oferecidos;
- lacuna(*gap*) 4: discrepância entre os serviços oferecidos e o que é comunicado ao cidadão; e
- lacuna(*gap*) 5: discrepância entre o que o cidadão espera receber e a percepção que ele tem dos serviços recebidos.

Para mensurar as 5 lacunas, propõe-se usar as dimensões da qualidade de serviço, quais sejam:

ISBN 978-85-61115-09-8

- aspectos tangíveis - equipamentos, pessoal envolvido e material de comunicação;
- confiabilidade - habilidade de prestar o serviço com exatidão;
- presteza - disposição em ajudar os cidadãos e fornecer o serviço com presteza e prontidão;
- segurança (garantia) - conhecimento dos funcionários e suas habilidades em demonstrar confiança; e
- empatia - grau de cuidado e atenção pessoal dispensado aos clientes (PARASURAMAN; ZEITHAML; BERRY, 1988).

A mensuração das lacunas resulta da comparação entre as percepções e/ou expectativas dos envolvidos. A qualidade dos serviços pode ser avaliada através de um levantamento entre aqueles que utilizam os serviços e os prestadores desses serviços. Com o objetivo de equiparar critérios entre tomadores e prestadores, pode utilizar-se uma escala Likert de cinco posições, variando de "concordo totalmente", com valor 5; "concordo na maior parte", valor 4; "não concordo nem discordo", valor 3; "discordo na maior parte", valor 2 e "discordo totalmente", com valor 1. Assim, o sujeito declara seu grau de concordância com afirmações predefinidas e é obtido o valor de cada sentença. Este servirá para calcular o grau de qualidade dos serviços prestados. Como resultado, tem-se:

- lacuna (*gap*) 1 – da expectativa: expectativa cidadão – expectativa governo
- lacuna (*gap*) 2 – da percepção: percepção governo – expectativa governo
- lacuna (*gap*) 3 – do desempenho: expectativa do cidadão – percepção do governo
- lacuna (*gap*) 4 da perspectiva: percepção do cidadão – percepção do governo
- lacuna (*gap*) 5 da satisfação: percepção do cidadão – expectativa do cidadão

Os valores identificados na escala Likert são tratados de forma a que possam ser comparadas as declarações entre cidadãos e governo. As diferenças poderão apresentar valores positivos, negativos ou neutros. Se o resultado for negativo, o serviço está abaixo do esperado, a experiência foi negativa. No caso de ser positivo, a experiência foi melhor do que era esperado. O resultado neutro significa inexistência de discrepâncias entre expectativas e percepções dos envolvidos.

No caso dos dados ligados, entregar um serviço conforme a expectativa do usuário seria relacionar (interligar) dados abertos até alcançar a informação desejada por ele e de acordo com as dimensões da qualidade da informação. Tal objetivo seria atingido no momento em que as lacunas fossem reduzidas ou eliminadas. Após definido o que os usuários precisam, as dimensões permitem embasar as decisões sobre o meio e a forma como elas podem ser oferecidas. As tabelas 1 e 2 apresentam os critérios de avaliação utilizados nesta oportunidade.

3 ESTUDO DE CASO

Para o propósito de este estudo foi escolhido o Instituto Brasileiro de Geografia e Estatística (IBGE). A seguir, é apresentada a instituição e analisa-se a forma de disponibilização dos seus dados.

ISBN 978-85-61115-09-8

3.1 Caracterização da Instituição

Em 1934 foi criado o Instituto Nacional de Estatística (INE), que iniciou suas atividades em 1936, no ano seguinte foi instituído o Conselho Brasileiro de Geografia, incorporado ao INE, que passou a ser chamado de Instituto Brasileiro de Geografia e Estatística (IBGE) (GONÇALVES, 1995).

Atualmente o IBGE é o principal provedor de dados e informações estatísticas do país. A instituição atende às necessidades dos mais diversos segmentos da sociedade civil, bem como de órgãos governamentais federais, estaduais e municipais. O IBGE é uma entidade da administração pública federal, que está vinculada ao Ministério do Planejamento, Orçamento e Gestão. Para que suas atividades possam cobrir todo o território nacional, o IBGE possui uma rede nacional de pesquisa e disseminação, composta por 27 unidades estaduais, 27 setores de documentação e disseminação de informações e 581 agências de coleta de dados nos principais municípios brasileiros (IBGE, 2014a)

Por meio de sua rede nacional de disseminação, com áreas de atendimento em todas as capitais e nas principais cidades, o IBGE oferece um dos maiores acervos especializados em informações estatísticas e geográficas do país. Este acervo constitui-se de publicações impressas e eletrônicas, como também de bases de dados (IBGE, 2014b). As estatísticas do mês de agosto de este ano revelaram um atendimento a 3.760.666 usuários, sendo que a grande maioria utiliza o portal de uma forma geral, fazendo consultas variadas e destaca-se que 17,5% acessam os dados e informações sobre cidades, no <cidades.ibge.gov.br> (IBGE, 2014c).

Um dos mais importantes canais de comunicação com o usuário é o Portal do IBGE na internet, onde são disponibilizados os resultados de pesquisas em páginas dinâmicas, com arquivos para *download* e banco de dados. Sendo o Sistema IBGE de Recuperação Automática (SIDRA) um dos principais e mais completos recursos de banco de dados estatísticos disponíveis no Brasil (ZANNOTO, 2011).

O SIDRA está disponível na *internet* gratuitamente, sendo acessível 24h por dia e 7 dias por semana, e conta com mais de 900 tabelas de dados totalizando 600 milhões de valores (JACON, 2006). O SIDRA possui recursos de busca por palavras-chave contidas nas tabelas ou na categorização dos metadados das pesquisas, por número de tabela e por tema ou seção. Também possui recursos para a personalização das formas de apresentação das tabelas para visualização ou geração em arquivo, assim como na geração de gráficos e cartogramas a partir dos dados pesquisados (ZANOTTO, 2011).

Dessa forma, o IBGE disponibiliza para os usuários em seu portal todos os dados estatísticos coletados, esses dados são armazenados na forma de tabelas, e permite que o usuário escolha quais dados serão agregados nas tabelas.

3.2 Proposta de Avaliação

A fim de ilustrar a proposta de avaliação do valor dos dados abertos ligados, foram utilizados os dados oferecidos pelo IBGE na sua página oficial na *web*. A aplicação resultou nas Tabelas 1 e 2.

A Tabela 1 apresenta as dimensões propostas no modelo, quais sejam: empatia, confiabilidade, segurança, presteza e tangibilidade. Para cada dimensão foram oferecidas duas afirmações a serem avaliadas seguindo a tabela Likert de 1 a 5. Os valores expressam expectativas e percepções sobre os dados ligados do IBGE, por parte de cidadãos e governo. Os dados registrados na tabela são meramente ilustrativos para elucidar a metodologia proposta.

ISBN 978-85-61115-09-8

Tabela 1 - Percepções e Expectativas

Dimensão	Afirmações	Governo/IBGE		Cidadão	
		Exp.	Perc.	Exp.	Perc.
Empatia	A ligação dos dados é semelhante àquela que eu usaria.	4	3	4	2
	Os dados são apresentados de uma forma interessante.	5	5	5	4
Confiabilidade	Os dados disponibilizados são atualizados.	5	5	5	4
	A proveniência dos dados é confiável.	5	5	5	5
Segurança	O acesso aos dados é liberado.	5	5	5	4
	Os links direcionam para o local esperado.	5	4	4	3
Presteza	O acesso aos dados é realizado de forma rápida.	4	3	4	3
	A apresentação dos dados facilita a navegação.	5	4	4	2
Tangibilidade	Os links representam o mundo real de forma adequada.	5	5	4	3
	Apresenta grande quantidade de dados de forma bem estruturada.	5	4	4	3

(concordo totalmente = 5; concordo na maior parte = 4; não concordo nem discordo = 3; discordo na maior parte = 2; discordo totalmente = 1)

Fonte: autores

A partir das afirmações, foram indicados os graus de concordância. Para as expectativas, o valor atribuído foi com relação ao que seria esperado, e as expectativas derivaram da navegação no portal do IBGE. Percebeu-se que muitos links se referem a documentos e não a dados, quando há a ligação entre dados, não direcionam a um URI específico.

Tabela 2- Divergências entre cidadãos e governo

Dimensão	Afirmações	GAP 1	GAP 2	GAP 3	GAP 4	GAP 5
Empatia	A ligação dos dados é semelhante àquela que eu usaria.	0	-1	1	-1	-2
	Os dados são apresentados de uma forma interessante.	0	0	0	-1	-1
Confiabilidade	Os dados disponibilizados são atualizados.	0	0	0	-1	-1
	A proveniência dos dados é confiável.	0	0	0	0	0
Segurança	O acesso aos dados é liberado.	0	0	0	-1	-1
	Os links direcionam para o local esperado.	-1	-1	0	-1	-1
Presteza	O acesso aos dados é realizado de forma rápida.	0	-1	1	0	-1
	A apresentação dos dados facilita a navegação.	-1	-1	0	-2	-2

ISBN 978-85-61115-09-8

Tangibilidade	Os links representam o mundo real de forma adequada.	-1	0	-1	-2	-1
	Apresenta grande quantidade de dados de forma bem estruturada.	-1	-1	0	-1	-1

Fonte: autores

As divergências encontradas constam na Tabela 2. As principais lacunas referem-se à Empatia (-2) e Presteza (-2) quando observado o gap 5, que é aquele que deve ser observado com cautela. Ele representa a expectativa que o usuário tinha antes de usar o portal e a impressão que ficou após. Neste caso, reflete a dificuldade de encontrar informações específicas, já que o portal se caracteriza por conter muita informação numa única página.

O *gap* 4 também apresenta valores -2. Esta lacuna traz a diferença entre o que é o oferecido e o que é comunicado sobre o que será oferecido. Isto também gera expectativas no usuário que, poderão não ser atendidas devido às questões de comunicação.

Evidenciar os *gaps* a partir das dimensões, com o uso de um instrumento de levantamento como o apresentado, contribuirá para a melhoria na prestação de serviços da instituição. Pois, tais evidências tem a capacidade de fundamentar mudanças na estrutura da página e na disponibilização dos mesmos.

4 Discussão

Os dados abertos do IBGE são muito importantes para que tanto instituições públicas quanto a sociedade em geral possam instruir-se sobre aspectos econômicos, demográficos, sociais e ambientais. Por ser a informação pública um direito ao cidadão, seus dados precisam ser oferecidos de forma a facilitar a interpretação correta do que eles representam, daí a importância da descrição dos dados.

Os dados do IBGE podem ser entendidos como dados abertos já que os dados estão disponíveis para todas as pessoas, é possível realizar o *download* e reutilizá-los, sem custo ou impedimentos.

Mas embora os dados do IBGE sejam atualizados frequentemente e em volumes significativos, eles poderiam ser melhor explorados se fossem disponibilizados em um padrão que fosse facilmente *linkado* com outras bases de dados. Isto aponta numa ampliação de possibilidades de uso ao serem relacionados a outras fontes, externas à instituição ou não.

No contexto do IBGE, os dados não se encontram no padrão RDF, impossibilitando a sua ligação com outras bases de dados da *web*, dessa forma podemos dizer que os dados abertos do IBGE são muito importantes para a descoberta de novos conhecimentos, porém como não possuem padronização, são de difícil reuso junto a outras bases de dados da *web*.

A adaptação do modelo ServQual permite evidenciar as 5 lacunas existentes entre os cidadãos e o governo a partir das percepções e expectativas de cada um deles. Ao construir o exemplo de aplicação, embora fictício, tentou-se reproduzir uma situação real que levou a evidenciar certas lacunas, principalmente quanto à Presteza, Empatia e Tangibilidade. As dimensões sobre Confiabilidade e Segurança foram consideradas com desempenho melhor.

Portanto, pode-se considerar que o IBGE possui dados abertos para visualização, mas existem inúmeras barreiras técnicas para que sejam reutilizados pela sociedade para criação de novos projetos e serviços. Vale salientar também que outra barreira ao uso é o conhecimento prévio que o usuário precisa ter para poder selecionar os elementos certos para formar as tabelas propostas pelo IBGE no portal. A montagem de tais tabelas demanda uma familiaridade com as variáveis utilizadas, seus critérios, além de saber utilizar planilhas eletrônicas para poder tirar proveito das mesmas.

ISBN 978-85-61115-09-8

Entretanto, já podem ser encontrados dados disponibilizados pelo IBGE noPortal Brasileiro de Dados Abertos. Também a participação do IBGE no INDA representam evidências na busca de agregar valor aos seus dados.

5 CONSIDERAÇÕES FINAIS

O IBGE destina uma verba significativa para o custeio de suas ações de levantamento dos dados. Para tal é necessário um planejamento minucioso, o envolvimento de muitos servidores, outras pessoas envolvidas de outras instituições, trabalhadores temporários, colaboradores, que demandam treinamento e instrumentos, assim como envolve a população em geral, como nos exemplos dos censos e levantamento de campanhas eleitorais.

Após esse esforço que reúne o trabalho de tantas pessoas, os dados podem retornar à população de outras formas, com outros benefícios extraídos, e não só aqueles que são oferecidos pelo próprio IBGE. Embora os relatórios realizados pelo IBGE sejam extremamente úteis, a ligação nos dados poderá resultar em ganho significativo ao relacionar informações sobre assuntos não oferecidos nos relatórios ou disponibilizados de outra maneira.

O governo brasileiro já conta com iniciativas para dados abertos como a Infraestrutura Nacional de Dados Abertos (INDA), Infraestrutura Nacional de Dados Espaciais (INDE-BR) e o Portal Brasileiro de Dados Abertos, o e-gov, mas encontra-se numa fase incipiente quanto a dados abertos ligados.

A política nacional sobre dados abertos, da qual o IBGE participa expondo seus dados no portal e do próprio INDA, todavia faz com que sejam perdidas inúmeras oportunidades de aproveitamento dos seus dados. Assim, uma forma de avaliar as estratégias que levem a melhorias nesse sentido, que indique o que disponibilizar à população, de que forma e agregação, poderá ser apontado mediante uma avaliação das lacunas da qualidade de serviço e da informação.

A partir das expectativas e necessidade dos cidadãos, definir as mudanças estruturais e tecnológicas que deverão ocorrer para que os dados da instituição e de outros oriundos em outras esferas públicas possam realmente trazer informação e conhecimento à sociedade. A proposta de uso dos 5 *gaps* foi escolhida pelo fato de considerar tanto o ambiente externo quanto o interno da organização. Cabe salientar que, se existe o benefício decorrente de ligar dados abertos, também há oportunidades perdidas pela falta de tais ligações, o que merece um estudo futuro, desde o ponto de vista econômico.

REFERÊNCIAS

ALBRECHT, Karl. Vocação para os Serviços. **HSM Management**, pp. 47 – 54, especial, 2000.

ALI, Syed T.; SIVARAMAN, Vijay; OSTRY, Diethelm; JHA, Sanjay Jha. Securing data provenance in body area networks using lightweight wireless link fingerprints. Proceedings of the 3rd international workshop on Trustworthy embedded devices, Pages 65-72, ACM New York, NY, USA 2013.

BERNERS LEE, Tim.**Linked Data**. Documento online. 2009 (última . Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acessoem: 28 ago. 2014.

BERNERS-LEE, Tim; O'HARA, Kieron. The read–write Linked Data Web. **Philosophical Transaction of The Royal Society**, v. 371, n. 1987, 2013. Disponívelem: <http://rsta.royalsocietypublishing.org/content/371/1987/20120513>. Acesso em: 12 set. 2014.

ISBN 978-85-61115-09-8

BIZER, Christian; HEATH, Talis; BERNERS-LEE, Tim. Linked Data: the story so far.**International Journal on Semantic Web and Information Systems**, v. 5, n. 3, pp. 1-22,2009.Disponível em: <http://go.galegroup.com/ps/i.do?id=GALE%7CA209477051&v=2.1&u=capes&it=r&p=AONE&sw=w&asid=bba599ce7e0043b3ca80e22dba03c18b>. Acesso em: 28 ago. 2014.
BIZER, Christian; HEATH, Tom; IDEHEN, Kingsley; BERNERS-LEE, Tim. Linked Data on the Web (LDOW2008). **Proceedings** of the 17th international conference on World Wide Web, Beijing, 2008. Disponível em: <http://dl.acm.org/citation.cfm?id=1367760>

DINIZ, Vagner. Como Conseguir Dados Governamentais Abertos. Congresso Consad de Gestão Pública, 3, Brasilia, 15-17 mar. 2010. **Anais** …

FERRARAM, Alfio; NIKOLOV, Andriy; SCHARFLE, François. Data Linking for the Semantic Web. Information Science Reference, Hershey PA, 2013

GONÇALVES, Jayci de Mattos Medeira. **IBGE: um retrato histórico**. Rio de Janeiro: IBGE, 1995. 61p. (Documentos para disseminação. Memória institucional, n. 5).

GRAY, Jonathan; et al. **Manual dos Dados Abertos**: Governo. Traduzido e adaptado de opendatamanual.org. Original revisado em jan. 2011. Disponível em: <http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>. Acessoem: 28 ago. 2014.

IBGE. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **O IBGE**. Rio de Janeiro: IBGE, 2014a. Disponível em: <http://www.ibge.gov.br/home/disseminacao/eventos/missao/instituicao.shtm> Acesso em: 28 ago. 2014.

IBGE. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Disseminação.**Rio de Janeiro: IBGE, 2014b.Disponível em: <http://www.ibge.gov.br/home/disseminacao/eventos/missao/disseminacao.shtm>
Acesso em: 28 ago. 2014.

IBGE. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Estatísticas do Site.** Rio de Janeiro: IBGE, 2014c.Disponível em: <http://www.ibge.gov.br/home/disseminacao/online/estatisticas/>Acesso em: 28 ago. 2014.

JACON, Maria C. P. Banco de Dados do IBGE: desafios tecnológicos. In.: Conferência Nacional de Estatística, 5, Rio de Janeiro, 2006; Conferência Nacional de Geografia e Cartografia, 4, Rio de Janeiro, 2006. **Anais** … Rio de Janeiro: IBGE, 2006.

JONES, Patricia M. Collaborative Knowledge Management, Social Networks, and Organizational Learning. 2001. Disponível em:
<http://human-factors.arc.nasa.gov/publications/collab_know_paper.pdf> Acesso 28 ago. 2014

KANE, G. G.; ALAVI, M. Information technology and organizational learning: an investigation of exploration and exploitation processes. **Organization Science**, n. 18, p. 796–812, 2007.
KASCHESKY, Michael; SELMI, LUIGI. 7R Data Value Framework for Open Data in Practice: Fusepool. Future Internet 2014. Disponível em: <www.mdpi.com/1999-5903/6/3/556/pdf> Acesso em: 28 ago 2014

NONAKA, Ikujiro; TAKEUCHI, Hirotaka. **Criação de Conhecimento na Empresa**. Rio de Janeiro: Esevier, 1997.

OPEN DEFINITION. Site oficial do projeto. Disponível em: <www.opendefinition.org>. Acesso em: 28 ago. 2014.

ISBN 978-85-61115-09-8

OPEN KNOWLEDGE FOUNDATION. **Open Data Handbook Documentation**. 2012. Disponível em: <http://opendatahandbook.org/pdf/OpenDataHandbook.pdf> Acesso em: 28 ago. 2014.

PARASURAMAN, A.; ZEITHAML, Valarie; BERRY, Leonard.A Conceptual Model of Service Quality and Its Implications for Future Research.**Journal of Marketing**, v. 49, pp. 41-50, 1985.Disponívelem: <http://faculty.mu.edu.sa/public/uploads/1360593395.8791service%20marketing70.pdf>. Acesso em: 02 set. 2014.

PARASURAMAN, A.; ZEITHAML, Valarie; BERRY, Leonard. Servqual: a multiple-item scale for measuring consumer perceptions of service quality. **Journal of Retailing**, v. 64, n. 1, pp. 12-40, spring, 1988. Disponívelem: <http://areas.kenan-flagler.unc.edu/Marketing/FacultyStaff/zeithaml/Selected%20Publications/SERVQUAL-%20A%20Multiple-Item%20Scale%20for%20Measuring%20Consumer%20Perceptions%20of%20Service%20Quality.pdf>. Acesso em: 02 set. 2014.

PORTAL BRASILEIRO DE DADOS ABERTOS. Site oficial. Disponível em: <http://dados.gov.br/>. Acesso em: 12 set. 2014.

WALLIS, J.C; ROLANDO, E; BORGMAN, C. L. If We Share Data, will anyone use them? **Data Sharing and Reuse in the Long Tail of Science and Technology**. 2013. Disponível em: <http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0067332&representation=PDF>. Acesso em: 28 ago 2014.

ZANOTTO, Regina Sônia. **Informação estatística oficial produzida pelo IBGE: apropriação pela comunidade científica brasileira no período de 2001 a 2009**. Dissertação de Mestrado Programa de Pós-Graduação em Comunicação. Universidade Federal do Rio Grande do Sul, Faculdade de Biblioteconomia e Comunicação, Porto Alegre, 2011.

ISBN 978-85-61115-09-8

PROPOSTA DE UM OBSERVATÓRIO DE SOFTWARE NO BRASIL COM RECURSOS DA WEB SEMÂNTICA

Márcio Martins Da Silva
marcio.martins@copasa.com.br
Luiz Cláudio Gomes Maia
luiz.maia@fumec.br
Fernando Silva Parreiras
fernando.parreiras@fumec.br

Resumo

Diversas soluções têm sido propostas para tentar melhorar a pesquisa e compartilhamento de dados na Web. Entretanto, a tecnologia conhecida como Web Semântica e Dados Ligados se destaca por propor uma padronização e integração universal de publicação de dados utilizando para este fim a estrutura atual da Web. Nesta linha, o objetivo deste trabalho é utilizar os conceitos da Web Semântica para propor um arcabouço conceitual para construção de um repositório com dados de fontes na Web sobre a indústria de software no Brasil. Esse repositório servirá como um observatório para que empresas multinacionais de software tenham uma visão integrada do cenário desta indústria no Brasil, auxiliando as mesmas nos processos de planejamento estratégico e tomada de decisão para definição de políticas setoriais em relação às suas subsidiárias no país.

Palavras-chave: observatório de software; dados abertos; web semântica

Abstract

Several solutions have been proposed to improve search and data sharing on the Web. However, the technology known as Semantic Web and Linked Open Data stand out for proposing a standardized and universal integration of data publication using for this purpose the current structure of the Web. With this approach, the aim of this work is to use the concepts of the Semantic Web to propose a conceptual framework for building a repository of data from sources on the Web about the software industry in Brazil. This repository will serve as an observatory for Multinational software companies to give them an integrated view of the scenery of this industry in Brazil, assisting in the process of strategic planning and decision making for defining sectorial policies regarding their subsidiaries in the country.

Key Words: observatory of software; linked open data; semantic web

ISBN 978-85-61115-09-8

Introdução

Devido à sua capacidade de melhorar processos, garantir eficiência e ganhos de produtividade, as atividades de software e serviços de Tecnologia da Informação (TI) são tidas como estratégicas pelo governo brasileiro. Dessa forma, políticas públicas diversas tratam de criar as condições necessárias como capacitação de pessoal, incentivos à inovação, à exportação e incentivos fiscais diversos, para o fortalecimento e a consolidação da indústria local (SOFTEX, 2012).

Durante os últimos vinte anos, a indústria brasileira de software e serviços de Tecnologia da Informação (TI) vem crescendo a taxas elevadas. Trata-se de uma indústria altamente diversificada, com produtos, soluções e serviços maduros e de alta complexidade, testados e aprovados pelo mercado e direcionados para os mais variados setores e segmentos econômicos: finanças, telecomunicações, gestão empresarial, saúde, educação, entretenimento, agronegócios e outros (SOFTEX, 2012).

Existe hoje um volume significativo de dados sobre a indústria de software em diversos sites brasileiros, principalmente nos governamentais. Entretanto, a falta de padronização na publicação destes dados, em grande parte disponibilizada em formatos proprietários ou apenas para visualização, cria uma série de obstáculos à reutilização dos mesmos. Além disso, sites com informações sobre assuntos similares muitas vezes não possuem nenhum tipo de compartilhamento ou conexão entre eles, produzindo-se com isto verdadeiras ilhas de informações.

O uso da tecnologia Web Semântica proporciona a integração de dados em fontes heterogêneas, potencializando a descoberta de novos conhecimentos. A WEB Semântica é uma extensão da WEB original, em que programas de aplicação conhecidos por agentes conseguem percorrer as páginas e desempenhar tarefas sofisticadas para os usuários, ao invés de se limitarem simplesmente a apresentá-los. A WEB desenvolveu-se originalmente como um meio para publicação de documentos, mas a tendência é que as páginas sejam codificadas de modo que seus conteúdos possam ser processados de forma automática (PATRÍCIO, 2010).

Em decorrência do que foi exposto, percebe-se a importância da construção de uma estrutura que auxilie empresas transnacionais de software nos processos de planejamento estratégico em território brasileiro. Essa estrutura deverá ser capaz de se sobrepor à heterogeneidade das fontes de informações, para que possa oferecer uma visão integrada dos dados a seus usuários.

Nesse sentido, este trabalho tem o intuito de apresentar um arcabouço conceitual para construção de um Observatório de Software, utilizando fundamentos da WEB Semântica, para que Empresas Multinacionais (EMNs) deste setor tenham acesso a um repositório contendo informações relevantes sobre esta indústria nos diversos municípios do Brasil, para que as mesmas tenham uma visão adequada da situação desta indústria no país, e possam decidir sobre políticas de investimento, pesquisa e desenvolvimento no país.

ISBN 978-85-61115-09-8

A Figura 1 apresenta um resumo que permite hierarquizar fatores essenciais para atração de investimentos internacionais em pesquisa e desenvolvimento, segundo opiniões de autores. Identifica-se aqueles recorrentes com maior frequência, indicando a atratividade de um como polo de investimentos e negócios.

Figura 1 - Principais fatores para atração de Multinacionais

Fator	Autores
Tamanho de Mercado	Mengistu (2009) Galina *et al*. (2011) Brain (2011) Stal e Campanário (2011) Negri e Laplane (2009) Bortoluzzo *et al.* (2012) Amal *et al*. (2007) Zucoloto (2012)
Oferta de mão de obra qualificada	Mengistu (2009) Galina *et al*. (2011) Brain (2011) Kinda (2008) Bortoluzzo *et al.* (2012) Negri e Laplane (2009) Zucoloto (2012)
Infraestrutura Básica	Mengistu (2009) Ageyman-Duah (2012) Galina *et al*. (2011) Brain (2011) Stal e Campanário (2011) Kinda (2008) Bortoluzzo *et al* .(2012) Zucoloto (2012)
Custo de mão de obra	Galina *et al.* (2011) Stal e Campanário (2011) Kinda (2008) Bortoluzzo *et al.* (2012) Amal *et al.* (2007)
Universidades e Institutos de Pesquisa	Galina *et al.* (2011) Stal e Campanário (2011) Negri e Laplane (2009) Zucoloto (2012) Silva *et al.* (2012)
Crescimento de Mercado	Zucoloto (2009) Galina *et al*. (2011) (Brain, 2011)
Parques Tecnológicos	Zucoloto (2012) Luo (2008) Silva *et al.* (2012)
IDH	Brain (2011) Amal *et al*. (2007)
Formação de Recursos Humanos	Zucoloto (2009) Donaubauer *et al.* (2013)
Incubadoras de Empresas	Zucoloto (2012) Luo (2008)
Incentivo (fiscal e subsídios)	Galina *et al*. (2011) Negri e Laplane (2009) Zucoloto (2012)

Fonte: Própria

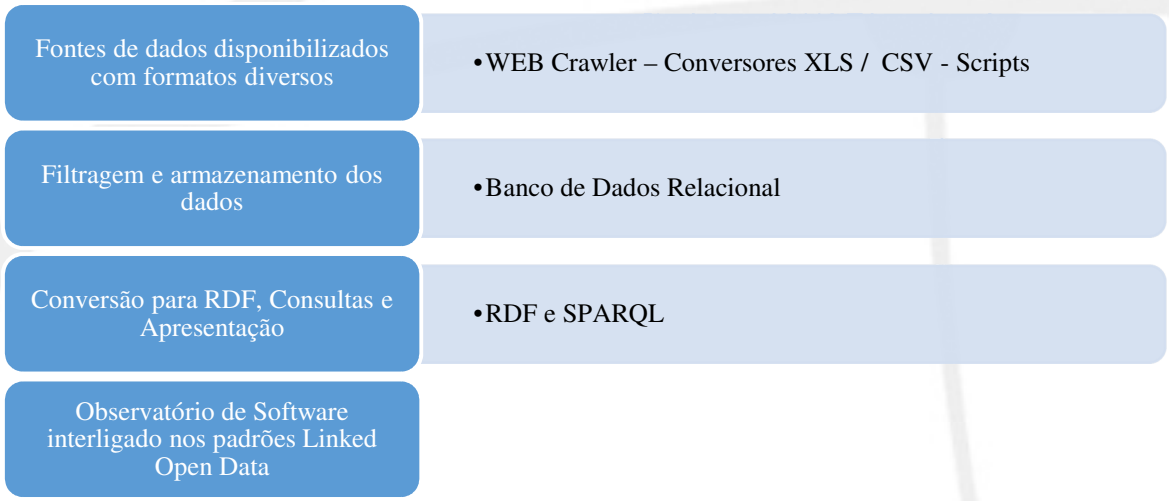
ISBN 978-85-61115-09-8

1 Proposta de Um Observatório de Software: Arcabouço Conceitual

O projeto tem como foco a execução de um conjunto de tarefas para a criação, armazenamento e publicação de dados seguindo o padrão de LinkedOpenData.

A figura 2 ilustra a arquitetura geral do projeto compõe de três camadas. A primeira camada é composta por fontes com dados estruturados e não estruturados, dispersos na WEB e que possuam informações que respondam às questões relacionadas aos fatores de atração de empresas de P&D no setor de desenvolvimento de software.

Figura 2 – Arquitetura geral do projeto



Fonte: Própria

O fluxo para criação dos Dados Ligados se inicia com a identificação e seleção de informações que serão armazenadas no repositório proveniente de diferentes fontes, mapeamento destes dados para representações semânticas, e, como última etapa, a publicação dos mesmos.

A camada seguinte consiste de procedimentos para extração dos dados das fontes escolhidas e inserção em tabelas do banco de dados relacional, onde serão classificados e filtrados, com exclusão de informações que não estejam relacionadas aos objetivos do projeto.

Na terceira e última camada, os dados em tabelas do banco de dados são mapeados para representações semânticas no modelo RDF, que é o padrão utilizado para publicar dados na WEB Semântica. Esta camada é composta por uma plataforma que permite consultas por *queries* SPARQL, e *interface* para apresentação HTML.

Assim como os sistemas de bancos de dados fazem uso do SQL para consultar registros em bases de dados, SPARQL é a linguagem de consulta para recuperação de informações contidas em arquivos RDF.

Usuários do sistema poderão utilizar *queries* já escritas para os fatores de atratividade levantados ou desenvolver novas consultas para explorar toda a base criada para o projeto.

ISBN 978-85-61115-09-8

2 Implementação

O Banco de Dados *MySql* foi escolhido para armazenar e organizar os dados trazidos das fontes na *web*.

A geração do *dataset* para o observatório de software baseou-se em quatro linhas de ação: a obtenção dos dados nas diversas fontes na WEB, inserção e filtragem dos dados no Banco de Dados *MySql*, e conversão dos mesmos para o modelo RDF.

A última etapa consistiu no desenvolvimento de *queries* SPARQLpara consulta à base de dados e respostas às questões de pesquisa para os fatores considerados vitais à atração de empresas transnacionais de *software.*

Depois de identificadas as fontes na WEBpara os diversos fatores de atratividade já descritos, foram desenvolvidos programas para busca, extração e armazenamento de informações no banco de dados.

Para extrair os dados das fontes, códigos de busca conhecidos como *crawlers* foram escritos em linguagem *Python,* com o objetivo de navegar páginas WEB de uma forma automatizada e ordenada, e criar cópias das mesmas para pós-processamento por programas *screenscrapers.*

Páginas WEBsão normalmenteformatadas em linguagem HTMLpara finsde visualização. Os programas *screenscrapers,* que também foram escritos para este projeto em linguagam *Python*, capturaram as saídas destinadas a um utilizador humano, retirando códigos de atributos e posicionamento, e armazenando somente os dados de interesse em arquivos no formato CSV (*comma-separated values*), que foram então importados para tabelas do banco de dados.

Para criação de um *dataset* para o Observatório de *Software*, foram utilizados dados disponibilizados pelo Ministério do Trabalho e Emprego (MTE), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), Associação Nacional de Entidades Promotoras de Empreendimentos Inovadores (ANPROTEC), Programa das Nações Unidas para o Desenvolvimento (PNUD), Instituto Brasileiro de Geografia e Estatística (IBGE) e da Associação Brasileira das Instituições de Pesquisa Tecnológica e Inovação (ABIPTI).

A Figura 4 identifica as fontes de dados utilizadas para o trabalho, e o formato em que as informações estão disponibilizadas para acesso.

Figura 4 - Fontes, dados e formatos utilizados no trabalho

Fonte	Dado	Formato
Ministério do Trabalho	RAIS	Skydrive
IBGE	Dados de municípios brasileiros	HTML
CAPES	Mestrados e Doutorados	XLS
INEP	Universidades e cursos	XLS
ANPROTEC	Incubadoras e Parques Tecnológicos	XLS
PNUD	Índice de Desenvolvimento Humano	XLS
ABIPTI	Institutos de Pesquisa	HTML

Fonte: Própria

ISBN 978-85-61115-09-8

Para conversão dos dados inseridos na base de dados para o modelo RDF,foi utilizada a plataforma D2RQ (**e***sse é um sistema de código aberto, desenvolvido com o objetivo de permitir acesso a banco de dados como grafos virtuais RDF. Desta forma, não é necessária a replicação das informações das tabelas do banco de dados em vocabulários RDF).*

Um componente do D2RQ é o *generating-mapping*. Essa é uma linguagem declarativa que possibilita o mapeamento nos formatos RDF/XML, NOTATION3 (N3) e N3 TRIPLE., a partir do esquema do banco de dados. Gera-se, desta forma, relacionamentos entre tabelas e colunas do banco para classes e propriedades de ontologias.

Uma interface SPARQL ou SPARQL *endpoint* permite consultas à base de dados, com características de *linked data*, ou seja, as URIs retornadas pelas queries podem acessadas diretamente desta interface.

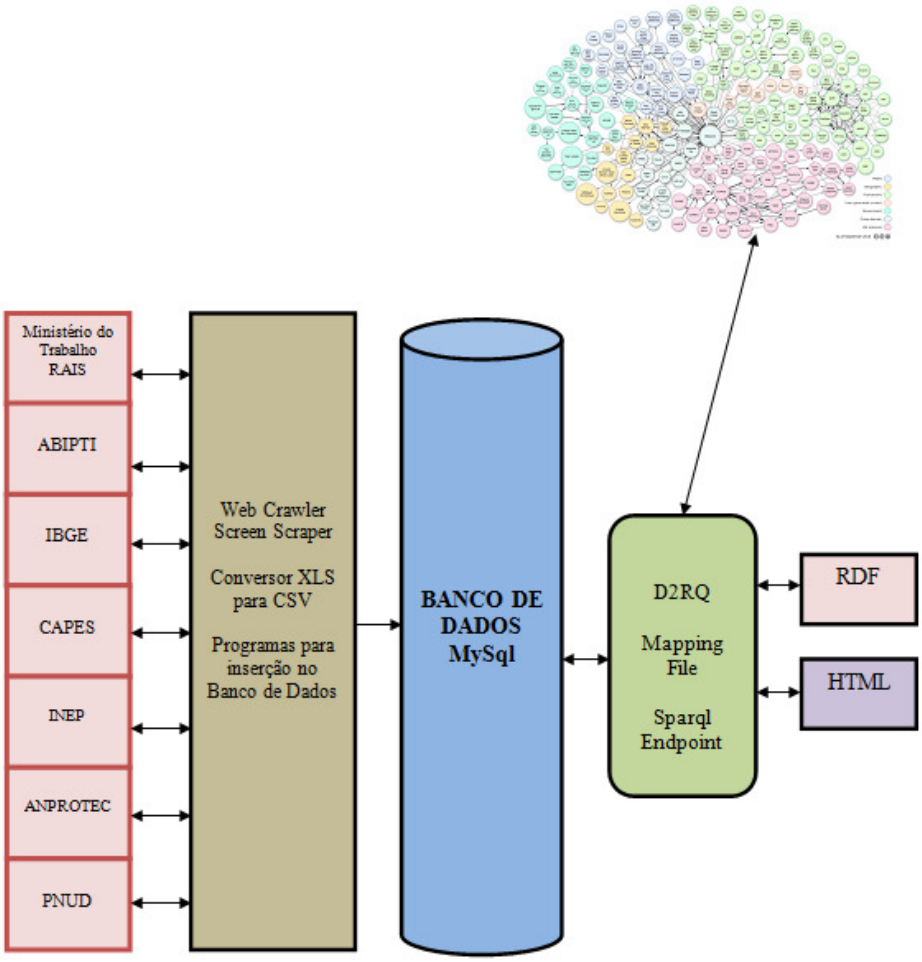
Por meio do arquivo de mapeamento, queries em linguagem SPARQL são convertidas dinamicamente para queries SQL, eliminando a necessidade da replicação dos dados em plataformas externas ao banco de dados, também conhecidas como *triple store*. Sendo assim, inserções, deleções e atualizações nas tabelas mapeadas, são visualizadas imediatamente por queries SPARQL, da mesma forma que acontece com queries sql executadas diretamente na base de dados.

Usando-se uma linguagem declarativa, define-se um mapeamento entre o esquema relacional do banco de dados e o vocabulário RDF. Com base neste mapeamento, o servidor D2R publica uma visão dos Dados Ligados, permitindo também consulta através de *queries* SPARQL (CYGANIAK; BIZER, 2006).

A Figura 3 ilustra a arquitetura geral do projeto, identificando os diferentes módulos e conexões entre os mesmos.

ISBN 978-85-61115-09-8

Figura 3 - Arquitetura geral do projeto



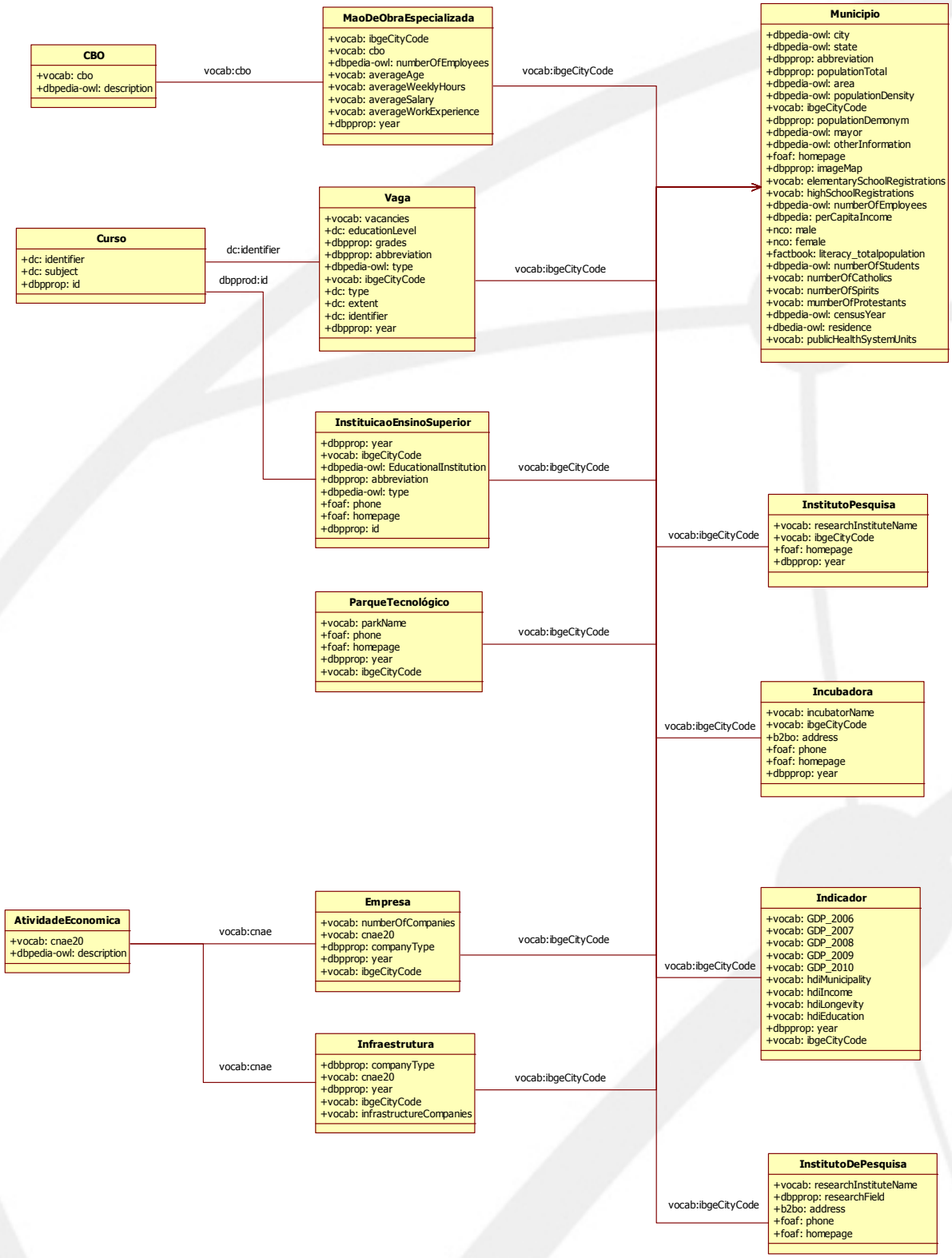
Fonte: Própria

Para o projeto do Observatório os termos de vocabulários bem conhecidos foram reutilizados. Nos casos em que houve a impossibilidade de reutilização, novos termos foram criados com nomenclatura em inglês, para que possam ser reaproveitados por desenvolvedores de aplicações em dados ligados.

O diagrama de classes para o Observatório de Software está ilustrado na Figura 5:

ISBN 978-85-61115-09-8

Figura 5 - Diagrama de Classes completo para Observatório de Software



Fonte: Própria

ISBN 978-85-61115-09-8

4 Validação

Nesta seção serão respondidas questões de pesquisa relacionadas ao fator mão de obra especializada, com respectivas consultas SPARQL e telas de resultados gerados no SPARQL *endpoint* da plataforma D2r-server.

Para todos os demais fatores levantados foram desenvolvidas consultas SPARQL para realização de prova de conceito e validação do projeto.

A figura a seguir mostra uma consulta aos códigos e ocupações brasileiras do cadastro do Ministério do Trabalho e Emprego, através de uma *query* SPARQL.

Figura 6 - Query SPARQL para códigos e descrições de ocupações

```
SELECT DISTINCT ?descricao ?cbo  WHERE {

?s vocab:cbo ?cbo .

?s dbpedia-owl:description ?descricao.

}

order by ?descricao
```

Fonte: Própria

O resultado dessa consulta é uma relação com todas as ocupações armazenadas na tabela.

Figura 7 - Resultado da *query* com todas as ocupações existentes

SPARQL results:

descricao	cbo
"Abatedor"	"848505"
"Acabador de Calcados"	"764305"
"Acabador de Embalagens (Flexiveis e Cartotecnicas)"	"766305"
"Acabador de Superficies de Concreto"	"716105"
"Acougueiro"	"848510"
"Acrobata"	"376205"
"Adestrador de Animais"	"623005"
"Administrador"	"252105"
"Administrador de Banco de Dados"	"212305"
"Administrador de Edificios"	"510110"
"Administrador de Fundos e Carteiras de Investimento"	"252505"
"Administrador de Redes"	"212310"
"Administrador de Sistemas Operacionais"	"212315"
"Administrador em Segurança da Informação"	"212320"
"Advogado"	"241005"
"Advogado (Areas Especiais)"	"241030"
"Advogado (Direito Civil)"	"241015"
"Advogado (Direito Penal)"	"241025"
"Advogado (Direito Publico)"	"241020"
"Advogado (Direito do Trabalho)"	"241035"
"Advogado da Uniao"	"241205"

Fonte: Própria

A seguinte consulta retorna todos os cargos considerados como amostra, para o município de Porto Alegre, RS. A seleção foi realizada utilizando-se os códigos CBO.

ISBN 978-85-61115-09-8

Figura 8 - *Query* SPARQL para diversas ocupações

```
SELECT DISTINCT ?cidade ?estado ?mobra ?descricao  WHERE {

?municipio dbpedia-owl:city "Porto-Alegre" .
?municipio dbpprop:abbreviation "RS" .
?municipio dbpedia-owl:city  ?cidade .
?municipio dbpedia-owl:state ?estado .
?municipio vocab:ibgeCityCode ?codmun .
?mobra vocab:ibgeCityCode ?codmun .
?mobra vocab:cbo ?cbo .
FILTER (regex(?cbo, "212315") || regex(?cbo, "212405") || regex(?cbo, "212415") || regex(?cbo, "212420") ||
regex(?cbo, "212410") || regex(?cbo, "212320") || regex(?cbo, "317110") || regex(?cbo, "317105")) .

?mobra vocab:averageAge  ?idademedia .
?s vocab:cbo ?cbo .
?s dbpedia-owl:description ?descricao .
}
order by ?descricao
```

Fonte: Própria

O resultado da consulta é ilustrado abaixo:

Figura 9 - Resultado da pesquisa para município

SPARQL results:

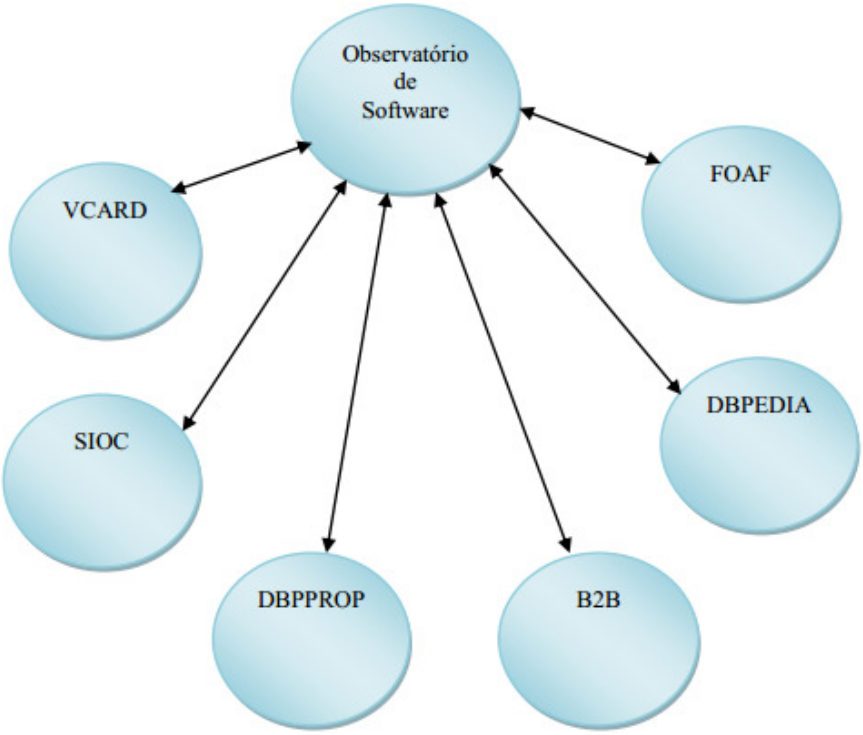
cidade	estado	mobra	descricao
"PORTO-ALEGRE"	"Rio-Grande-do-Sul"	<http://localhost:2020/resource/mao_de_obra/7938>	"Administrador de Sistemas Operacionais"
"PORTO-ALEGRE"	"Rio-Grande-do-Sul"	<http://localhost:2020/resource/mao_de_obra/8407>	"Administrador em Segurança da Informação"
"PORTO-ALEGRE"	"Rio-Grande-do-Sul"	<http://localhost:2020/resource/mao_de_obra/8101>	"Analista de Desenvolvimento de Sistemas"
"PORTO-ALEGRE"	"Rio-Grande-do-Sul"	<http://localhost:2020/resource/mao_de_obra/8370>	"Analista de Redes e de Comunicacao de Dados"
"PORTO-ALEGRE"	"Rio-Grande-do-Sul"	<http://localhost:2020/resource/mao_de_obra/8171>	"Analista de Sistemas de Automacao"
"PORTO-ALEGRE"	"Rio-Grande-do-Sul"	<http://localhost:2020/resource/mao_de_obra/8280>	"Analista de Suporte Computacional"
"PORTO-ALEGRE"	"Rio-Grande-do-Sul"	<http://localhost:2020/resource/mao_de_obra/8685>	"Programador de Internet"
"PORTO-ALEGRE"	"Rio-Grande-do-Sul"	<http://localhost:2020/resource/mao_de_obra/8497>	"Programador de Sistemas de Informacao"

Fonte: Própria

Com todo o modelo alimentado, a proposta de ligação com as fontes públicas de vocabulários está ilustrada na figura 10:

ISBN 978-85-61115-09-8

Figura 10 - Ligação entre o *dataset* Observatório de Software e fontes de públicas de vocabulários



Fonte: Própria

5 Trabalhos Relacionados

Nesta seção descreve-se dois projetos semelhantes ao proposto, o primeiro é o sistema www.ligadonospoliticos.com.br desenvolvido na Universidade Federal de Juiz de Fora (UFJF), é composto por uma base de dados com informações sobre políticos brasileiros, através de coleta de informações de diversos sites na WEB. Os dados coletados são agrupados e transformados em formato de dados ligados. Esses dados foram estruturados, utilizando para isto o modelo RDF, reutilizando-se vocabulários descritos em esquemas de metadados conhecidos como *Friend of a Friend* (FOAF), SKOS e outros, e criados outros termos para aqueles que não constavam nos esquemas de metadados utilizados.

O segundo trabalho relacionado com este artigo é o Observatório da Associação para Promoção da Excelência do *Software* Brasileiro (SOFTEX). O Observatório SOFTEX é a unidade de estudos e pesquisas da SOFTEX, uma organização da sociedade civil de interesse público, sediada em Campinas, São Paulo. O objetivo deste observatório é coletar, organizar, analisar e difundir dados e informações sobre as atividades de *software* e serviços de Tecnologia da Informação (TI) realizadas no Brasil (SOFTEX, 2012).

O observatório de *software* da Softex não possui a característica proposta neste artigo de atualização dinâmica dos dados no repositório em intervalos determinados, fazendo com que quaisquer mudanças de informações nas fontes sejam modificadas na base de dados do sistema.

ISBN 978-85-61115-09-8

6 Conclusão

A proposta deste estudo mostrou que é viável a construção de um Observatório de *Software* dentro do contexto da WEB Semântica. A base semântica criada possibilita consultas a informações sobre características diversas dos mais de 5500 municípios brasileiros, com algumas consultas já estabelecidas, assim como a possibilidade de desenvolvimento de novas consultas, além da possibilidade de interligação com outras bases na WEB.

O Observatório foi construído utilizando dados reais, levantados junto a ministérios e órgãos reconhecidos oficialmente, propiciando uma avaliação realista da indústria de *software* no país.

O Observatório de *Software* também pode ajudar a modificar a imagem do Brasil perante aos investidores internacionais, já que segundo a consultoria Ernest & Young (EY, 2012), mais de 55% de executivos entrevistados destas empresas, se limitam a reconhecer São Paulo capital como a região mais atraente para investimentos, seguido por Rio de Janeiro, com 26%.

A divulgação de características de outros municípios fora desse eixo pode contribuir para atração de investimentos estrangeiros para outras regiões, o que seria decisivo para um desenvolvimento do país rumo a uma forma mais equitativa.

Dentre as dificuldades para implementação prática do sistema, podemos citar a reduzida quantidade de informações em formatos abertos, mesmo entre os órgãos governamentais, o que dificulta a reutilização de informações de páginas de ministérios e outros órgãos públicos no país.

Apesar do sistema dinâmico, a ideia inicial de construção de um sistema completamente automático, sem a necessidade de atualização em intervalos determinados, não pode ser concretizada, já que algumas fontes de dados não permitem que agentes de *software* realizem a extração de informações em suas páginas, como já citado o caso do Ministério do Trabalho e Emprego (MTE), onde somente transferências manuais são possíveis.

A descoberta de termos para composição da base de dados semântica foi outro desafio encontrado. Existem serviços na WEB que auxiliam na descoberta de vocabulários, já que o reuso de termos já modelados está entre as melhores práticas para dados ligados.

No entanto, esses serviços ainda são bastante limitados, já que não fornecem uma indicação clara de melhores vocabulários para reutilização, o que pode levar à criação de termos já existentes para o domínio em questão, que prejudicam a interoperabilidade entre aplicações.

Atento às demandas contemporâneas para a área de Tecnologia da Informação (TI), o projeto prevê a ampla disponibilização do seu sistema para o acesso universal. Coloca-se apenas momentaneamente restrito pelas limitações de capacidade de processamento e armazenamento do equipamento utilizado no desenvolvimento do projeto. Este entrave, porém, pode ser sanado desde que se obtenha transferência dos dados e programas para um

ISBN 978-85-61115-09-8

servidor com maior capacidade de recursos, ou a disponibilização destes em serviços de *cloud computing*, pelo comportamento escalável e dinâmico deste tipo de infraestrutura.

Esperamos, com este trabalho, contribuir para que os conceitos da WEB semântica sejam amplamente utilizados no desenvolvimento de novas aplicações, para que o Brasil passe a ter, no futuro, um papel de vanguarda no cenário internacional nesta nova geração da WEB, o que pode significar ótimas oportunidades de negócios, além de potencializar o crescimento científico e tecnológico do país.

Referências

AGEYMAN-DUAH, R. P. **Nation branding as a tool for the increase of foreign direct investment.**[S.l.:S.n.], 2012. Disponível em: <http://air.ashesi.edu.gh/bitstream/handle/123456789/17/done%20-%20RACHEAL%20POKUAH%20AGYEAN%20-%20DUAH.pdf?sequence=1>. Acesso em: 21 jan. 2013.

AMAL, M. et al. **Análise dos determinantes institucionais e regionais do investimento direto externo das pequenas e médias empresas: um estudo do caso da região sul do brasil.**[S.l.:S.n.], 2007. Disponível em: <http://www.periodicos.ufsc.br/index.php/economia/article/download/2276/1929>. Acessoem: 14 nov. 2012.

BORTOLUZZO, M. M. et al.**Alocação do Investimento Direto Externo entre estadosbrasileiros.**[S.l.:S.n.], 2012. Disponível em:<http://en.insper.edu.br/sites/default/files/2012_wpe269.pdf>. Acesso em: 5 jan. 2013.

BRAIN. **Atratividade do Brasil como polo internacional de investimentos e negócios.**[S.l.:S.n.], 2011. Disponível em: <http://brainbrasil.org/relatorios brain/lan/br/id/e84bd268004eb9fdab2e47709ba17a91>. Acesso em: 22 dez. 2012.

CYGANIAK, R. D2R SERVER A*ccessing Relational Databases as Virtual RDF Graphs.* [S.l.:S.n.], 2012. Disponível em: <http://d2rq.org/d2r-server>. Acesso em: 03 fev. 2013.

DONAUBAUER. et al.**Does Aid for Education attract foreign investors ?An EmpiricalAnalysis for Latin America.**[S.l.:S.n.], 2013. Disponível em: <http://www.econstor.eu/handle/10419/67340>. Acesso em: 5 dez. 2013.

GALINA, S. V. R. et al.**Fatores de Atração de Atividades de Pesquisa e Desenvolvimento(P&D): um survey das filiais de empresas multinacionais instaladas no Brasil.**[S.l.:S.n.], 2010. Disponível em: <http://www.anpec.org.br/encontro2010/inscricao/arquivos/0008454b20bd7b90c91a295f2d0ce2bd376.doc.>. Acesso em: 18 dez. 2012.

KINDA, T. **Investment Climate and FDI in Developing Countries: Firm-Level Evidence.** [S.l.:S.n.], 2008. Disponível em: <http://economics.ca/2008/papers/0339.pdf>. Acesso em: 15 dez. 2012.

LIGADO NOS POLÍTICOS. **Dados governamentais abertos.** [S.l.:S.n.], 2011. Disponível em: <http://ligadonospoliticos.com.br>. Acesso em: 12 dez. 2012.

LUO, Y. **China: Current trends in pharmaceutical drug discovery.**[S.l.:S.n.], 2008. Disponível em: <http://www.gnipharma.com/japanese/news/download/20080417.pdf>. Acesso em: 17 dez. 2012.

ISBN 978-85-61115-09-8

NEGRI, F. D.; LAPLANE, *M.* **Fatores locacionais e o investimento estrangeiro em p&d:Evidências para o brasil, argentina e México.**[S.l.:S.n.], 2009. Disponível em: http://www.anpec.org.br/encontro2009/inscricao.on/arquivos 000877becf3800b21d2f07d9710d2ede1c.doc>. Acesso em: 19 dez. 2012.
MENGISTU, A. A. **The Roles of Human Capital and Physical Infrastructure on FDIInflow: Empirical Evidence from East Asia and Sub Saharan Africa.**[S.l.:S.n.], 2009. Disponível em: <http://www.csae.ox.ac.uk/conferences/2009-EDiA/papers/122 Mengistu.pdf>. Acesso em: 2 dez. 2012.

PATRÍCIO, H. S. **A Europeana e a agregação de metadados na web: análise dosesquemas ESE/EDM e da aplicação de standards da web semântica a dados debibliotecas.**[S.l.:S.n.], 2010. Disponível em:<http://bad.pt/publicacoes/index.php/congressosbad/article/view/458>. Acesso em 19 nov.2012.

SOFTEX. **Software e Serviços de TI: A indústria brasileira em perspectiva** [S.l:S.n.],2012. Disponível em: <http://www.mbi.com.br/mbi/biblioteca/papers/2012-06-softex industria-software-ti-perspectiva-volume-2/2012-Observatorio-Softex-Industria-Brasileira Software-Servicos-TI-em-perspectiva-Versao-Completa-Portugues.pdf>. Acesso em 07 jan.2013.

STAL, E.; CAMPANÁRIO, M. A. **Inovação em subsidiárias de empresas multinacionais:a aplicação do paradigma eclético de Dunning em países emergentes.**[S.l.:S.n.], 2011.Disponível em: <http://www.scielo.br/scielo.php?pid=S141323112011000200010&script=sci_arttext>. Acesso em: 21 nov. 2012.

SILVA, M. F. O. et al. **Incentivos para a implantação de centros de P&D internacionaisno Brasil**. [S.l.:S.n.], 2012. Disponível em: <http://www.bndes.gov.br/SiteBNDES/export/sites/default/bndes_pt/Galerias/Arquivos/conh cimento/bnset/set3601.pdf> Acesso em: 15 jan 2013.

ZUCOLOTO, G. F. **Origem de capital e acesso aos incentivos fiscais e financeiros àinovação no Brasil.**[S.l.:S.n.], 2012. Disponível em: <http://www.ipea.gov.br/portal/images//stories/PDFs/TDs/td_1753.pdf>. Acesso em: 12 out. 2012.

ISBN 978-85-61115-09-8

Simulation and Structural

Analysis ofThematic Social Networks

Pablo Lucas, Luiz Antônio Moro Palazzo

University of Essex, Faculty of Social Sciences, England
Federal University of Santa Catarina, EGC, Brazil

plucas@essex.ac.uk , *luiz.palazzo@ufsc.br*

Resumo
O conceito de Redes Sociais Temáticas (RST) encontra-se em formação e há um amplo interesse sobre em que circunstâncias e com que grau de efetividade estas poderiam ser empregadas em processos de produção de conhecimento compartilhado. Particularmente em ambientes produtores de *linked open data*. Para investigar estas questões, é necessário aprofundar o estudo de RST. Aborda-se aqui a análise estrutural de suas conexões e a simulação de RST como um sistemas multiagente (SMA). O conceito de RST é definido e uma proposta para a simulação parametrizada de múltiplas RST é formalizada com um pseudo algoritmo para a geração sistemática de RSTs. A proposta inclui a descrição e modelagem estatística usando técnicas da análise de redes sociais (ARS). O objetivo é melhor entender as estruturas e propriedades das RST que podem ser produzidas pelos usuários finais. Conclusões são apresentadas, junto com o potencial de futuras pesquisas da aplicação de SMA e ARS ao conceito de RST.

Palavras-chave: redes sociais temáticas, modelagem baseada em agentes, análise de redes

Abstract
The concept of Thematic Social Networks (TSN) has been defined and is currently being implemented. There is institutional interest to understand how to best deploy TSNs, and under which circumstances the process of knowledge production and sharing can thrive through them. To investigate these two questions, an in-depth analysis of TSNs is proposed in terms of: (a) developing a simulation of TSNs and (b) statistically modelling their network structure. The former will be tackled by systematically generating TSNs via an agent-based simulation (ABM). The latter will be tackled with social network analysis (SNA) techniques to provide a statistical description and modelling of the links within and between TSNs. This is aimed at anticipating the structures and properties that could be created by end users, as the hereby-proposed tool is aimed at integrating ABM and SNA capabilities to better understand TSNs.

Keywords: thematic social networks, agent-based modelling, network analysis

1 Introduction

Collaborative online environments can be deployed in different knowledge areas, such as scientific research, educational settings and socio-economic settings. All these activities produce knowledge, documents, media and other outcomes that can be made available through the concept of linked open data. Such content, however, is often not structurally organised and formally described by the involved individuals –and for this reason, the sharing potential may remain unrealised. For that, it is necessary to make datasets openly available and also enable these to be directly used online, via

ISBN 978-85-61115-09-8

computational standards, without any pre-processing requirement [Skjæveland et al 2013, EC 2013].

The concept of a Thematic Social Network (henceforth TSN) is aimed at managing data access and linkage through a systematic computational description of the processes of knowledge production within specialised teams. Members shall engage in such environments, where a semi-automated process of resource description makes available both the structure and content of each TSNs. A pilot TSN is currently being implemented in Brazil and therefore the system is not yet available to end-users.

This paper proposes ways to anticipate the likely structures, within and between TSNs, through the development of an agent-based simulation for the generation of TSN structures that are driven by heterogeneous entities. The assumptions for generating structures, between and within TSNs, will be derived from the pilot TSN and then tested systematically. Further to that, social Network Analysis techniques will be applied to statistically describe and model the structure of links between and within TSNs. Our aim is to discuss the research design that tackle these research questions:

(a) How can simulations facilitate testing assumptions about the generation of TSNs?

(b) How can analytical techniques be used to understand the structure of linked TSNs?

(c) How can the overall structure of TSNs facilitate the sharing of linked open data?

2 Thematic Social Networks

Thematic Social Networks (TSN) can be defined according to the twelve attributes: (1) small-scale, (2) private, (3) closed, (4) own a theme that is (5) planned and (6) contains metadata, which is (7) formally developed via (8) collaborative activities stemming from (9) its participants, resulting in a (10) computational object that can be (11) exported and (12) shared on the Web [Palazzo, 2014]. Therefore, a TSN:

(1) is a computational entity that represents a close-knit and small team. Within such networked structure, one may reorganise participants within subgroups.

(2) is private, both in terms of its own infrastructure and knowledge that is collaboratively developed in teams. Some members only access; others control.

(3) is closed in the sense of a team having its own identity, which is physically manifested offline. E.g. a research group, a sports club or a student cohort.

(4) owns a theme, which is simultaneously the object of study and also the computational product stemming from the collaborative tasks of its members.

(5) has a development plan, that is intended to facilitate the achievement of collaborative objectives and milestones, which goes beyond social activities.

ISBN 978-85-61115-09-8

(6) has its own theme computationally represented, content-wise, using the Resource Description Framework and Attributes (henceforth RDF / RDFa).

(7) has a development scheme in which every participant has a clear role and, if needed, also assigned to different subgroups within the same network theme.

(8) has a collaborative work environment, with tools and resources that can be shared amongst its network members. E.g. blogs, archives, messaging, agenda.

(9) has members that are embedded in a network, as it evolves thematically.

(10) has a computational object derived from the collective knowledge of its members, which includes their relationships, interactions and outcomes.

(11) is wholly exportable, possibly into different metadata formats, so that one is able to restore its full state (content and structure) in a new environment.

(12) can store multiple versions, some of which may be published online, allowing third-party applications to make use of linked data technologies.

Each of the twelve attributes is required for a TSN to exist. Hence the architecture of a TSN is composed of the following layers: (a) information technology infrastructure, (b) the social network environment and (c) the collaborative environment of the TSN.

TSNs have the potential to facilitate collaborating online and sharing of linked data, as the concept is built upon computational standards that enable automated description of structure and content. Hence this is applicable to the relationships within and between TSNs, providing a way to help standardising the description of network datasets. That is important due to privacy concerns and the general difficulty to collect network data. For example, science and education projects could be continuously shared, each with their own development stage and rules of accessibility between TSNs. Such a new way of collaboration is intended to encourage much greater participation and transparency between those individuals involved in the development of knowledge that is online.

3 An Agent-Based Model approach to simulate the formation of TSNs

Models are central to science and appear in numerous formats. These range from physical representations (e.g. architectural scale models) all the way through to abstract mathematical specifications (e.g. models of theorem generalisations). The usefulness of models is closely associated with type and area of application. Some models provide theoretical representations, by means of implementing specific hypotheses that ought to deductively contribute to either corroborate or falsify theories through experiments. Other models are intended to represent a phenomenon for inductive explorations on how assumptions and processes interplay. Despite the diversity of scientific models, these are aimed at being somehow useful, matching at least one of following objectives: interpreting or describing data (e.g. statistically or qualitatively), testing interventions (e.g. deductive experimentation or inductive exploration), synthesising

ISBN 978-85-61115-09-8

studies, forecasting (e.g. econometrics) and/or idea elucidation (e.g. demonstration) [Sarkar, 2013, Heckman and Vytlacil, 2007]. Therefore the common feature amongst scientific models is to: "provide a simplified representation, of what is known and relevant, that is somehow helpful to further understand, describe or explore the real phenomena" [Geweke, 2005].

An agent-based model (henceforth ABM) is a computational method that enables a researcher to carry out experiments via simulations in which heterogeneous, autonomous agents interact. This computational social science approach is about building models as computer programs to represent aspects of a social phenomenon, such as a TSN. With an ABM one can aim to better understand the dynamics of a phenomenon through the processes of model building and data analysis. The entities in an ABM can processes input data (i.e. a fully specified scenario) to generate output based on the non-linear interaction between the implemented behavioural assumptions. The program itself represents a phenomenon, which may include a number of different actors (i.e. individuals, organizations. firms, nations). These are programmed with reactive and pro-active behaviours, which may be adapted (e.g. evolve) at runtime according to the model specifications. A crucial feature of an ABM, which sets itself apart from other computational approaches, is the interaction between multiple heterogeneous entities.

One can best understand the role of an ABM by thinking about these terms of usefulness, rather than truthfulness. I.e. the model may not be absolutely true[40] regarding the social phenomenon, yet it may still be useful to some degree regarding the real system. ABMs consist of individual entities with local rules and dynamics (i.e. a bottom- up), which generate macro outcomes (i.e. based on their aggregate) and simultaneously allow different levels of abstraction (i.e. incorporation of both quantitative and qualitative evidence) [Epstein, 2006, Gilbert, 2007]. These are formal (i.e. computational) representations of a phenomenon with the explicit goals of: (I) describing the target system more precisely than with non-discursive languages and (II) allowing experimentation via simulations [Moss and Edmonds, 2005].

Therefore the implementation of a TSN an ABM framework will open up opportunities to:

(a) systematically mediate the acquisition of insights via computational test of assumptions;
(b) clarification of potential issues in the TSN model specification via rigorous replications;
(c) suggest further developments, regarding TSNs, which would otherwise not be known.

The intention of this research proposal is to implement assumptions about a pilot TSN, into an ABM, and thus enable the exploration of trajectories both at the macro (i.e. system-wide) and micro (i.e. individual entity) levels. This research design allows carrying out systematic computational experiments that test the interplay between the specific endogenous variables without exogenous interferences. The ABM will then consist of would-be scenarios, based on justified assumptions, which allow the exploration and understanding of how to best deploy TSNs. The proposed contribution thus is to provide insights about the structure and dynamics of TSNs that would be unattainable without an agent-based simulation model. Given the TSN specifications in section 2, the pseudo-algorithm in Figure 1 below provides the proposed design on how the ABM will be implemented, focused on the simulation of TSN links.

[40] A model is never, representation-wise, fully accurate and correct regarding a social phenomenon.

ISBN 978-85-61115-09-8

Figure 1 – the ABM pseudo-algorithm for TSN link creation

```
REPEAT
        SET the maximum number of TSNs                          (minimum 2)
        SET the number of heterogeneous agents per TSN          (systematically from 10 to 40)
SET a theme for each TSN                                        (randomly between 2 and 5)
        SET teams within TSNs                                   (randomly between 3 and 6)
        SET connections between and within TSNs                 (to be defined based on the pilot)
UNTIL the maximum number of TSNs has been reached
```

The ABM shall be developed to deal with the structural link formation of TSNs, as depicted in Figure 2, and assumptions of individual behaviour will be based on insights acquired from deploying the TSN pilot. Then the analysis of the simulated, and empirical, TSN structures will be done using the set of techniques presented in the next section.

4 Application of Social Network Analysis to TSN

The analysis of structures within and between TSNs can be done with the set of techniques developed for Social Network Analysis (henceforth SNA). This is to be focused on elements of Layer C, depicted in Figure 2 below, and property (10) mentioned in section 2. That is, the object derived from the collective knowledge of its members, in particular the relationships and interactions between the TSN members. Therefore, SNA has the potential for allowing:

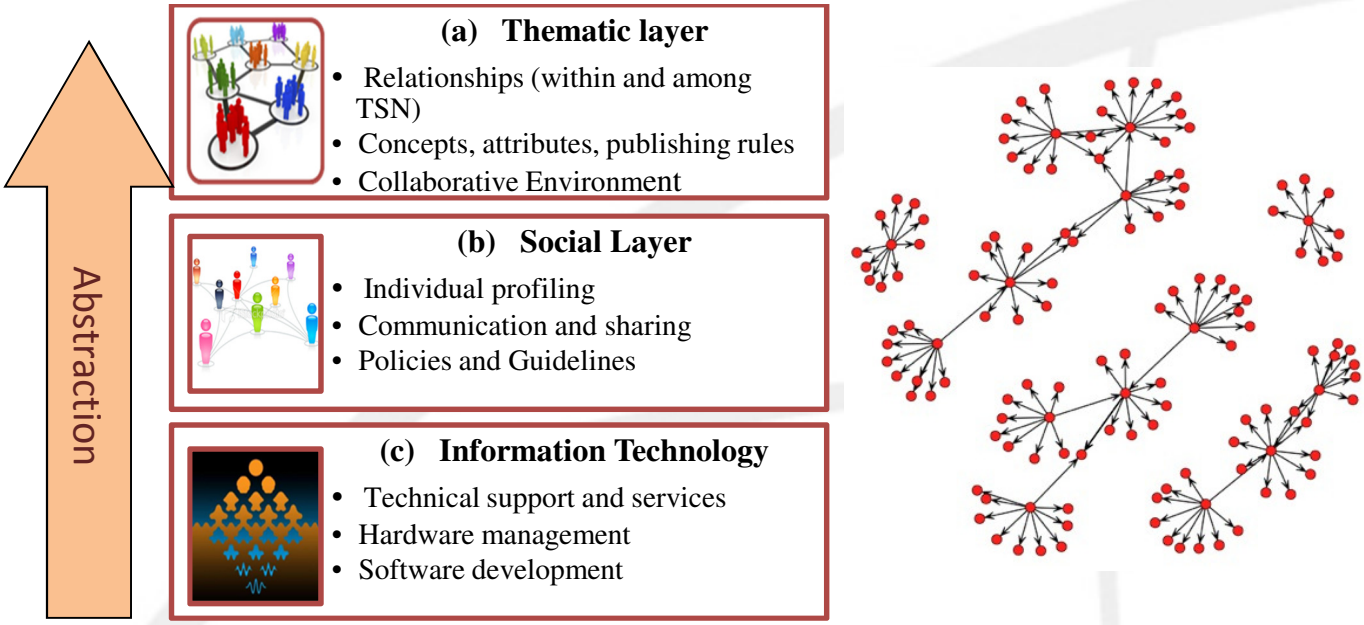
(a) statistical description of the structure and inference of the likely processes, social (bottom-up) and institutionalised (top-down), that lead to patterns within and/or between TSNs;

(b) cross-sectional and longitudinal analyses, including the structural dynamics (e.g. centrality measures such as closeness, betweenness, eigenvector and degrees, such as in and out) plus node and link-level covariates (e.g. attributes of members and TSNs themselves).

Besides providing the descriptive statistics, the workhorse for the modelling proposal is the class of models entitled Exponential Random Graphs (henceforth ERGM), which include extensions for both one-shot and multiple-wave data analyses. An ERGM approach allows the identification of: informal networks (i.e. referring to structures that are endogenous and self-organising, such as reciprocity, homophily, hierarchy and transitivity) plus attribute-related networks (i.e. referring to effects that are driven by behaviour or roles) [Robins et al 2007].

ISBN 978-85-61115-09-8

Figure 2 – TSN layers and a sample network structure



ERGMs have a probabilistic approach to understand the links between a set of nodes, which can provide a useful statistical platform to investigate hypotheses regarding network structure. Such ongoing developments in network model specification and estimation can allow one to probe into networks using observablecharacteristics (i.e. structural patterns and node attributes). Regularities in social networks are generated by complex behaviour that can be statistically described and modelled, with some degree of error. The aim is to describe the properties of a network at a given time and draw inferences about how a TSN may evolve over time. Solely qualitative interpretations regarding network evolution pose difficulties to the understand which factor(s) contribute most to observed structural effects, so a quantitative approach can facilitate the assessment of potentially conflicting explanations regarding the network structure.

Statistically simulated networks contain, by definition, properties assigned at its specification. On the other hand, empirical networks result from social processes that do not follow a formal specification and thus require the estimation of parameters from the data itself. In absence of this information, the ABM outcome will thus be based on the non-linear interactions between the heterogeneous entities belonging to different TSNs. This is an important difference as otherwise one would not be able to explore, before end users actually work with TSNs, the structural properties of connections between and within the participants of TSNs.

5 Final Remarks

This research proposalpresents a design to understand the structure of TSN populations generated through an ABM approach and analysed with SNA techniques. The strategy relies on (a) using the ABM flexibility to systematically create heterogeneous TSN configurations plus (b) applying SNA to describe and model the TSN structures. The proposal set out is relevant to understand the likely TSN structures before their deployment to end-users.

ISBN 978-85-61115-09-8

Bibliography

European Commission (EC). Interoperability Solutions for Public Administrations (2013). How Linked Data is transforming Government. Accessed on 09/09/2014: http://goo.gl/P5IaGj

Epstein, J. M. (2006). Generative Social Science: Studies in Agent-Based Computational Modeling (Princeton Studies in Complexity). Princeton University Press.

Geweke, J. (2005). Contemporary Bayesian Econometrics and Statistics (Wiley Series in Probability and Statistics). Wiley-Interscience.

Gilbert, N. (2007). Agent-Based Models (Quantitative Applications in the Social Sciences). Sage, ISBN: 9781412949644, annotated edition.

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation, Handbook of Econometrics, volume 6 of Handbook of Econometrics, Elsevier.

Moss, S. and Edmonds, B. (2005). Sociology and simulation: Statistical and qualitative cross-validation. American Journal of Sociology, 110:1095–1131.

Palazzo, L. (2014). Redes Sociais Temáticas. UFSC/EGC Research Report, Florianópolis, Brazil.

Robins, G.; Pattison, P.; Kalish, Y.; Lusher, D. (2007). An introduction to exponential random graph models for social networks. Social Networks 29: 173–191.

Sarkar, S. (2013). Scientific Models in Philosophy of Science: An Encyclopaedia. Routledge, 2nd edition.

Skjæveland, M.G., Lian, E.H., Horrocks, I. (2013). Publishing the Norwegian Petroleum Directorate's FactPages as Semantic Web Data. International Semantic Web Conference.

ISBN 978-85-61115-09-8

ISBN 978-85-61115-09-8

Índice Remissivo

ISBN 978-85-61115-09-8