Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Semantic Enrichment of Web Data for the Provision of an Unified Data Repository of Brazilian Missing Persons

Versão completa através de:
coord.computacao@ice.ufjf.br

Jorão Gomes Junior

JUIZ DE FORA
JULHO, 2019

Semantic Enrichment of Web Data for the Provision of an Unified Data Repository of Brazilian Missing Persons

Versão completa através de:

coord.computacao@ice.ufjf.br

Jorão Gomes Junior

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Jairo Francisco de Souza

JUIZ DE FORA

JULHO, 2019

SEMANTIC ENRICHMENT OF WEB DATA FOR THE PROVISION OF AN UNIFIED DATA REPOSITORY OF BRAZILIAN MISSING PERSONS

Versão completa através de: coord.computacao@ice.ufjf.br

Jorão Gomes Junior

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Jairo Francisco de Souza
Dr. em Informática (PUC-RIO)

Victor Ströele de Andrade Menezes
Dr. em Engenharia de Sistemas e Computação (UFRJ)

Fabricio Martins Mendonça
Dr. em Ciência da Informação (UFMG)

JUIZ DE FORA

04 DE JULHO, 2019

Aos meus amigos e irmãos.

Aos pais, pelo apoio e sustento.

Resumo

As tecnologias de dados e comunicação estão se tornando intimamente ligadas à vida das pessoas. Por isso, é natural fazer uso de todo esse progresso para reduzir e resolver problemas sociais. Para que o governo e a sociedade tomem as decisões mais adequadas para lidar com o desaparecimento de civis, é necessário ter uma fonte de informação bem estruturada. Em vários países, é difícil acessar os dados do governo pois as informações estão dispersas, não conectadas e mal estruturadas. Assim, este trabalho apresenta um *framework* para coletar informações sobre o desaparecimento civil no Brasil por meio de técnicas como *Data Scraping* e *Linked Data*. O objetivo é disponibilizar uma centralização automática de dados desses casos individuais e incentivar o uso de padrões para a publicação de dados que são frequentemente ignorados pelas organizações, dificultando a análise e a tomada de decisão sobre esses dados.

O texto completo está em sigilo e pode ser requisitado através de: coord.computacao@ice.ufjf.br

Palavras-chave: Raspagem de dados, Dados ligados, Web Semântica, Pessoas desaparecidas.

Abstract

Communication and data technologies are becoming closely linked to people's lives. Therefore it is natural to make use of all this progress to reduce and solve social problems. In order for government and society to make the most appropriate decisions to deal with missing persons, it is necessary to have a well-structured source of information. In several countries, it is difficult to access government data, since information is dispersed, not connected and poorly structured. For this reason, this work presents a framework to gather information on missing persons in Brazil through techniques such as Data Scraping and Linked Data. The goal is to make available an automatic data centralization of these individual cases, and to encourage the use of standards for the publication of data that are frequently ignored by organizations, hindering analysis and decision making on data.

The full text is confidential and may be requested through: coord. computacao@ice.ufjf.br

Keywords: Data Scraping, Linked Data, Semantic Web, Missing Persons.

Agradecimentos

A todos os meus parentes, pelo encorajamento e apoio.

Ao professor Jairo, pela orientação, amizade e, principalmente, pela paciência, sem a qual este trabalho não se realizaria.

Aos professores do Departamento de Ciência da Computação, pelos seus ensinamentos, e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o nosso enriquecimento pessoal e profissional.

"So really the biggest mistake would be not to make that mistake, because then you'll go your whole life not knowing if something was a mistake or not".

Lily Aldrin (How I met your mother)

Contents

List of Figures

List of Tables

List of Abbreviations

HTML	Hypertext Markup Language
NGO	Non-governmental Organization
NLP	Natural Language Processing
OCR	Optical Character Recognition
XML	Extensible Markup Language

1 Expanded abstract

Over the years, it is possible to see the growth of the data volume published on the Web. Such growth is due to the fact of the existence of innumerable applications from which this data can be generated and used. However, interoperating among these data has always been a challenge and several publication patterns have been proposed as fundamentation of what is called the Semantic Web (BERNERS-LEE; HENDLER; LASSILA, 2001; BIZER et al., 2011) .

In the public sector, there is a great interest on the part of governments to have their information published openly. In Brazil, one of the main areas of government interest is the public security, which lacks specialized manpower and adequate investments to produce trusted national statistics (CERQUEIRA; LOBÃO; CARVALHO, 2005). In this scenario, it is difficult to find quality data in public security, as in the case of missing persons. There are some initiatives of the Brazilian government to encourage the localization of missing persons through the online disappearance registry, as in the Ministry of Justice[1]. However, the adhesion is low totaling 1206 registrations on this date, between missing and found persons.

Since each entity has its own set of rules to deal with the problem, the heterogeneity of information representation by each entity hinders the interoperability process between the data. The proposal to integrate these diverse data repositories in a structured and unified manner is a promising standardization alternative, since it allows a faster and more comprehensive exchange of information, allowing systems to use this data to aid in the dissemination process the occurrence of disappearance, the closure of cases, as well as facilitating the identification of duplicates and inconsistencies.

Therefore, this work aims, through a real-world application, to present the problems in the data structuring of the Web and to present a framework to improve the quality of making this data available. As an example, a case study will be presented with the manipulation of data date from missing persons cases.

[1] ⟨https://desaparecidos.mj.gov.br/⟩

1.1 Data extraction framework

To make the extraction and unification of the data for specific domain repositories easier, this are presents a framework for extraction, validation, grouping, and metadata attribution to these databases. The framework consists in a set of tasks for structuring data, semantic attribution and data normalization using approach *mediator & wrapper* (ÖZSU; VALDURIEZ, 2011).

The extraction process is performed using the data scraping technique. Extractors are defined from a single interface that has specific methods provided by the solution. These methods help in identifying the data layout on the page and assign each data to a property that represents its meaning in the vocabulary used by the repository. The definition of each collector is made using an XML document, which will be read and instantiated in the framework.

Each data generated by the extractor passes through a set of tasks defined by the developer. The purpose of the tasks is to enrich the data, extracting information, or performing some kind of normalization in the data. The proposed framework has a set of predefined tasks that can be used by any collector, such as the normalizer and the semantic annotation. In addition to normalization tasks, can be used tasks to enrich the data. From the use of metadata generation methods, one can infer data that is not necessarily explicit. In this context, Natural Language Processing (NLP) methods, Optical Character Recognition (OCR), or any other semantic aggregation tool or technique that the user application of the framework requires. Furthermore, to the available tasks, the user can include code in the PHP language itself, creating a class that implements an Task interface, where a task receives a collection of text and must also produce a collection of texts.

Finally, the data is stored in a central repository in RDF format. In this step, the data already has a structure of semantic representation of the attributes using vocabularies of the user's choice. When working with large volumes of data, there is a high probability of finding duplicated data. Therefore it is necessary to treat and verify these duplicate elements before inserting them in the database. To avoid duplicates, the user can define, for each extractor, the set of data that make up the identifier of the data. Before inserting a new record into the database, it is queried for its existence in the database using the

data set defined as identifier.

1.2 Unified data repository of missing persons

The motivation to apply in this scenario arose from the involvement with the Family Support Group of Missing Persons of the Federal University of Juiz de Fora, which aims to help families who have had missing relatives find their relatives. In this sense, the repository assists in the availability of data in a format that allows new applications to process this information more easily and allows a wide dissemination and visualization of the data of these people.

In this case, there are problems of duplication and completeness of data, where more than one source reproduces the same information and data that could be complementary from different websites, are disclosed separately. Still in this issue, another problem encountered is in the way that the HTML of outreach website are structured. Usually, information is presented in single blocks without any identification of what data is being presented or is presented in image format and PDF documents, making it difficult to collect and identify items.

To create the unified repository, an extensive Web search was conducted for website that contain data on Brazilian missing persons were founded 15 websites how does the dissemination of missing persons cases.

1.2.1 Extractors definition

Initially, the structures of the website and what types of data were offered by them were studied in order to make it possible to standardize the data collection. After the analysis, 19 different properties were requested during the registration of disappearance occurrences. To preserve the origin of these data, the source from which the data was extracted was also collected. For each source, a collector was instantiated in the framework, which generate data with properties of ontologies known in the literature.

1.2.2 Normalizer and definition of identifiers

For the scenario presented, all collectors use two simple normalization tasks: (1) the characters are transformed to lowercase and (2) all possible whitespaces are removed at the beginning and end of the extracted data.

In this scenario, a single missing person may be appearing on different websites. Thus, when entering the data of this same person, one must take the precautions to reduce duplicate data. For lack of a primary key in these data, identifiers in the collectors were the properties **name**, **disappearance date** and **disappearance place** of the individual.

1.2.3 Metadata attribution

In the sources that were selected, the information is presented in natural language text formats that do not have metadata for easy identification or are arranged in a key-value-like format, but also without the use of metadata. In some cases, the information is presented in image format or in PDF document. To address these issues, the following tasks were used, where the first two are programmed in the framework and the last one, being specific to this scenario, was developed separately and added as an extension of framework : Semantic Annotation, OCR and Gender Inference.

1.3 Results

After the extraction process, 11.242 records of missing persons were collected. Due to the lack of interaction between the sources, were found 743 duplicate records, treated by framework. Among these duplicates, 163 could be used to complement information from the registry, incomplete in the used sources.

After the framework process, can be seen that the data grew by approximately 60% for the birth date property, 40% for the age properties, 220% for the gender , 10% for states and 30% for cities. With these new data volumes, the overall growth of the unified repository was over 10%. This has made the base more homogeneous, with its consistent, correlated and centralized information improving the way it is viewed, disseminated and understood in cases of disappearance. Finally, the data generated by the framework

application is made available for use by other applications through a SPARQL endpoint and for querying others on a website.

1.4 Conclusion

This work presented a framework to support the extraction, structuring, standardization and semantic enrichment of Web data for the creation of specific domain repositories. As it has been proposed, framework allows the user to use existing tasks to improve data quality. In addition, framework extensions can be made to insert new tasks specific to the user's domain.

The framework was used within an real-world application that deals with data related to the problem of missing persons in Brazil. Fifteen website were identified as sources of information for the repository, which contained 11.242 missing records. After the collection and enrichment process, there was an increase of approximately 10% in the data present in the base. The repository created, as well as the collected data, is available to other users and applications. With this repository, it is hoped that this data can be used in the development of intelligent applications capable of assisting the competent authorities to develop public policies to combat missing persons problem.

Bibliography

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

BIZER, C. et al. Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, IGI Global, v. 1, n. 1, p. 205–227, 2011.

CERQUEIRA, D.; LOBÃO, W.; CARVALHO, A. X. d. *O jogo dos sete mitos e a miséria da segurança pública no Brasil*. Rio de Janeiro, RJ, Brasil: Instituto de Pesquisa Econômica Aplicada (Ipea), 2005.

ÖZSU, M. T.; VALDURIEZ, P. *Principles of distributed database systems*. [S.l.]: Springer Science & Business Media, 2011.