Quem@PUC - A tool to find researchers at PUC-Rio

**Mariana D. A. Salgueiro, Veronica dos Santos,
André L. C. Rêgo, Daniel S. Guimarães,
Edward H. Haeusler, Jefferson B. dos Santos,
Marcos V. Villas, Sérgio Lifschitz**

[1]Departamento de Informática, PUC-Rio

{msalgueiro, vdsantos, villas, sergio}@inf.puc-rio.br

{jsantos}@puc-rio.br

{andreregо, danielg}@aluno.puc-rio.br

Abstract. *Quem@PUC[1] is an Information Retrieval System available on the Web that allows searching for researchers and professors based on a keyword list of research related terms. It publicizes research and teaching activities from the PUC-Rio community to society in general. The idea is to integrate information from professors from administrative systems, courses offered, and researchers' Lattes CVs. Data sources are converted to RDF format using domain ontologies, then stored in a NoSQL database that supports native free-text indexing on triple objects. Search results include names, academic papers, teaching activities, and contact links.*

1. Introduction

How should one proceed when looking for experts on a particular subject, say, for a newspaper interview or to get involved in an R&D project? This situation often happens at every University, such as PUC-Rio and other well-known institutions. One possibility would be to perform a web search with the subject's keywords and add the expression "PUC-Rio". Another option would be to use Lattes platform, powered by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), to search by Subject (title or production keywords) inside the researches CVs. A third alternative would be to contact the "University Communication Area" and request a recommendation and contact details. Still, it is also difficult for them to find a specialist inside the university community.

Due to the research and academic information being spread across several sources, this project aims to bring them together in a single database and create a search engine capable of returning, through a keyword list, results in an organized way and with the proposal of doing this quickly and efficiently.

This paper describes an Information Retrieval System, named Quem@PUC[2], that allows the discovery of researchers and professors from PUC-Rio based on a keyword list. Through this tool, available on the Web, it is possible to publicize research activities from the PUC-Rio community to be used by society in general, considering the particular interest of other researchers as well as journalists. Search results include the names,

[1]https://bit.ly/3Cdejc1
[2]https://quemnapuc.biobd.inf.puc-rio.br/

academic papers, teaching activities, and contact links of those involved with the subject represented by the set of keywords entered.

2. Background

In most organizations, many information systems coexist and overlap in content and purpose, and at a university, like PUC-Rio, it is not different. In general, existing information systems were not designed to be integrated, and their corresponding databases were designed to meet a specific context. We may build an information integration system (IIS) to provide uniform access to a set of heterogeneous and autonomous information sources and extract insights from integrated information. Such systems abstract the complexity of locating and accessing the information silos and resolving conflicts between sources. Several approaches have been proposed, emphasizing solutions to structural and technical problems [Ziegler and Dittrich 2007].

IIS, where data from multiple sources and formats are stored in a single central repository, requires extraction, transformation, and load (ETL) processes. Tools specialized in creating, maintaining, and monitoring automated ETL processes do not require programming skills. However, the semantic information integration must guarantee that they will preserve those data interpretation aspects through the integration process, and only related data will be combined. Such requirement demands domain knowledge to discover, understand and represent information.

Ontologies define controlled vocabularies that uniquely identify a set of concepts to avoid ambiguity in their interpretation. Therefore, an ontology can also be seen as a data model that formally defines the relationships between concepts. Ontologies aim to establish an organized and standardized relationship between entities, enabling data contextualization extracted from different sources and facilitating interpreting them.

RDF (Resource Description Framework) is a W3C graph data model that allows data to be shared, reused, and described by ontologies. It has features that facilitate data integration and interoperability. Data and metadata can be represented through a triple schema (subject, predicate, object) and stored together in a TripleStore database. A TripleStore is a NoSQL graph database specialized in storing and manipulating RDF data through the SPARQL language. As it is a NoSQL data store, it also allows a flexible schema definition.

Information Retrieval Systems are designed to find digital objects, stored in extensive collections that satisfy user information need [Manning et al. 2008]. The query is usually specified in natural language through keywords representing the search intention. The search engine translates this query in its language model to return a list of digital objects. These digital objects can be documents, tables, graphics, videos, or images. Still, according to the object collection purpose and organization in more specific contexts, they may correspond to other resources like triples or subgraphs.

3. System Design and Technologies

Considering the system's purpose and the possibility of adding new data sources in the future, we have initially identified two non-functional requirements: (1) a schemaless data model, and (2) the support for loading files in different formats, such as XML, RDF, CSV,

and Excel spreadsheets. There is a variety of data sources and significant difficulty to establish a schematic model *a priori*. A data model without a rigid schema allows data to have variety in structure. Thus, we decided to store the integrated data in a NoSQL system with a flexible schema.

We have chosen RDF as the target data model of the ETL process. Standardized, known, and publicly available domain ontologies were used during the data integration transformation step to preserve the semantic of data extracted from different sources and facilitate data interpretation by the search tool. Resources (subjects and objects) must correspond to concepts present in the used ontologies as well as predicates to relationships and attributes. For example, we used *BIBO: Bibliographic Ontology Specification*[3], *BIO: A vocabulary for biographicalinformation*[4], *VIVO: An ontology for representing scholarship*[5], and *CCSO: Curriculum Course Syllabus Ontology*[6].

The Lattes platform, maintained by CNPq, is responsible for storing and making available Lattes Curriculum data of Brazilian researchers. CNPq provides a XML file for each Lattes Curriculum, as shown in figure 1. A XSLT script transformes each XML file into RDF using the selected ontologies as shown in figure 2. To transform other data sources such as administrative information about professors and academic courses offered at PUC-Rio, we use an ETL tool called Linked Pipes[7]. It is a lightweight tool specialized in triplifying, i. e., the process of transforming data from any structured or semi-structured format (relational, XML, CSV) into a triple format from the RDF graph model. Data source providers regularly dump the latest data, and we rerun the transformation pipelines and load the data into the repository.

```
<ARTIGO-PUBLICADO SEQUENCIA-PRODUCAO="154" ORDEM-IMPORTANCIA="">
    <DADOS-BASICOS-DO-ARTIGO NATUREZA="COMPLETO" TITULO-DO-ARTIGO="Nomenclatural novelties in
    Miconieae (Melastomataceae): new synonym and typifications" ANO-DO-ARTIGO="2020"
    PAIS-DE-PUBLICACAO="" IDIOMA="Português" MEIO-DE-DIVULGACAO="MEIO_DIGITAL"
    HOME-PAGE-DO-TRABALHO="[doi:10.11646/phytotaxa.443.2.5]" FLAG-RELEVANCIA="NAO" DOI="10.11646/
    phytotaxa.443.2.5" TITULO-DO-ARTIGO-INGLES="" FLAG-DIVULGACAO-CIENTIFICA="NAO"/>
    <DETALHAMENTO-DO-ARTIGO TITULO-DO-PERIODICO-OU-REVISTA="Phytotaxa (on-line)" ISSN="11793163"
    VOLUME="443" FASCICULO="" SERIE="2" PAGINA-INICIAL="179" PAGINA-FINAL="188"
    LOCAL-DE-PUBLICACAO=""/>
</ARTIGO-PUBLICADO>
```

Figura 1. An XML file of a given Lattes Curriculum

AllegroGraph was used as the database system for centralized storage of data originating from other systems. It is a multi-model NoSQL database (Document in JSON, JSON-LD, and Graph in RDF). We have chosen AllegroGraph because it allows the most significant number of triples (five million) per repository, compared to other available options in a free version. In addition, the database supports native free text (literals) indexing on triple objects. It is necessary to create free-text indexes where it is possible to specify which predicates should be considered, the stopwords to be removed, and the insensitive accent configuration.

[3]http://purl.org/ontology/bibo/
[4]http://purl.org/vocab/bio/0.1/
[5]https://duraspace.org/vivo/
[6]https://w3id.org/ccso/ccso#
[7]https://etl.linkedpipes.com/

```
<rdf:Description rdf:about="#P154">
  <rdf:type rdf:resource="http://purl.org/ontology/bibo/Article"/>
  <dc:title>Nomenclatural novelties in Miconieae (Melastomataceae): new synonym and typifications</dc:title>
  <dcterms:issued>2020</dcterms:issued>
  <dc:language>Português</dc:language>
  <foaf:homepage>[doi:10.11646/phytotaxa.443.2.5]</foaf:homepage>
  <bibo:doi>10.11646/phytotaxa.443.2.5</bibo:doi>
  <dcterms:isPartOf>
    <rdf:Description>
      <rdf:type rdf:resource="http://purl.org/ontology/bibo/Journal"/>
      <bibo:issn>11793163</bibo:issn>
      <dc:title>Phytotaxa (on-line)</dc:title>
    </rdf:Description>
  </dcterms:isPartOf>
  <bibo:pageStart>179</bibo:pageStart>
  <bibo:pageEnd>188</bibo:pageEnd>
  <bibo:volume>443</bibo:volume>
</rdf:Description>
```

Figura 2. An example of an XML file for a Lattes Curriculum Vitae converted to RDF with the help of ontologies

Subgraph matching queries were specified in the SPARQL language to retrieve resources of interest about researchers/professors. Keyword matching in queries was performed using the (*magic property*) *fti:match* operator, which allows the query to use the *freetext* index created. Due to repository size limitations, it was necessary to use a federation mechanism in order to search for the keyword list and return a result that involves all related repositories created in the database. This feature allows AllegroGraph to automatically distribute SPARQL queries among the involved repositories and combine the results transparently for the application.

Additionally, PostgreSQL DBMS was also used to record search activity to analyze user behavior and search terms. This log data allows statistics generation such as the most searched terms, most selected researchers/professors, etc. Also, in case of eventual problems, the data can provide information for quick fixes. Since log data is well defined and not subject to constant changes in the schema, we decided to adopt a Relational DBMS to meet this requirement.

The application was built using Python programming language in conjunction with the Flask microframework, allowing fast prototyping of web applications. The AllegroGraph Python API provides methods to create, query, and maintain RDF data and manage stored triples.

The entire system is divided into two servers: one containing the ETL tool (ETL Server), where all ETL pipelines are done, and the other is the application server (Web Server + DB Server), where the databases are hosted, and the search engine runs, managed with the Apache tool (see Figure 3).

4. Demonstration

Quem@PUC retrieves information given any list of keywords if any. The system shows results for exact matches given the input keywords. It returns items associated with professors/researchers that contain precisely the word(s) informed. We rank [Baeza-Yates and Ribeiro-Neto 1999] the results based on which professor/researcher has the most matches with the keyword input. In this pattern match operation, accented or up-

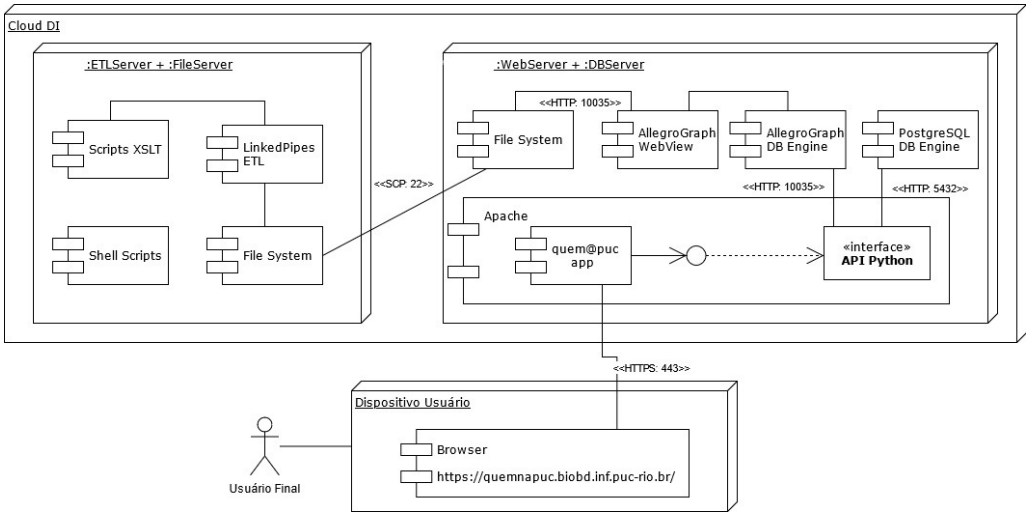
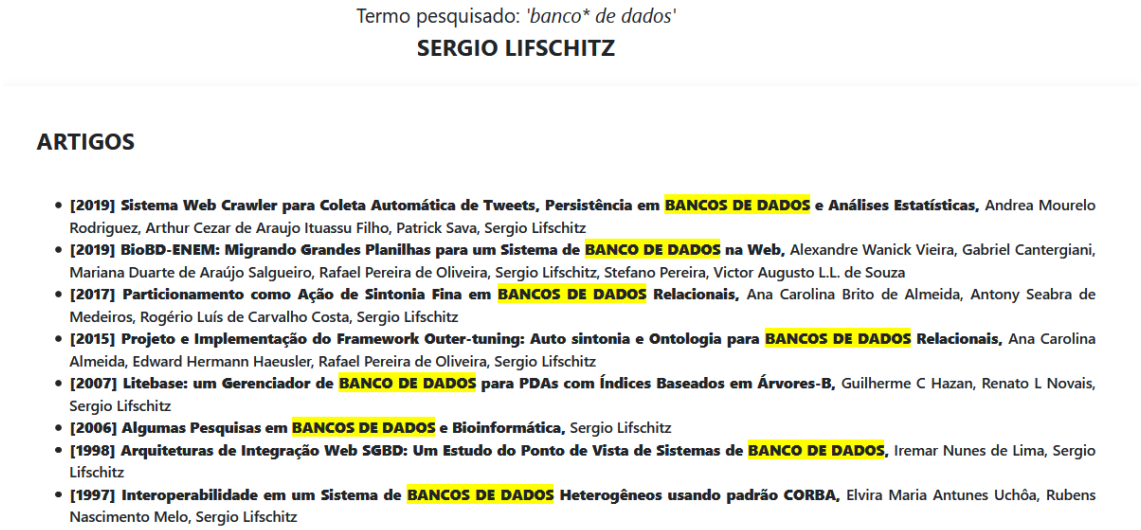


Figura 3. UML Deployment Diagram

percase characters are considered equivalent to unaccented or lowercase characters. The system performs an exact match with all the words input, in any order, that is, the logical operator to perform the matching is *AND*.

The system also allows an approximate search variation concerning the spelling of the words used. In this case, the user can make use of two *wildcards* character options. The question mark (?) matches any single character, and the asterisk (*) matches none to many characters at the associated position. In Figure 4, we present an example wildcard in the expression *banco* de dados*.


Termo pesquisado: '*banco* de dados*'

SERGIO LIFSCHITZ

ARTIGOS

- **[2019] Sistema Web Crawler para Coleta Automática de Tweets, Persistência em BANCOS DE DADOS e Análises Estatísticas,** Andrea Mourelo Rodriguez, Arthur Cezar de Araujo Ituassu Filho, Patrick Sava, Sergio Lifschitz
- **[2019] BioBD-ENEM: Migrando Grandes Planilhas para um Sistema de BANCO DE DADOS na Web,** Alexandre Wanick Vieira, Gabriel Cantergiani, Mariana Duarte de Araújo Salgueiro, Rafael Pereira de Oliveira, Sergio Lifschitz, Stefano Pereira, Victor Augusto L.L. de Souza
- **[2017] Particionamento como Ação de Sintonia Fina em BANCOS DE DADOS Relacionais,** Ana Carolina Brito de Almeida, Antony Seabra de Medeiros, Rogério Luís de Carvalho Costa, Sergio Lifschitz
- **[2015] Projeto e Implementação do Framework Outer-tuning: Auto sintonia e Ontologia para BANCOS DE DADOS Relacionais,** Ana Carolina Almeida, Edward Hermann Haeusler, Rafael Pereira de Oliveira, Sergio Lifschitz
- **[2007] Litebase: um Gerenciador de BANCO DE DADOS para PDAs com Índices Baseados em Árvores-B,** Guilherme C Hazan, Renato L Novais, Sergio Lifschitz
- **[2006] Algumas Pesquisas em BANCOS DE DADOS e Bioinformática,** Sergio Lifschitz
- **[1998] Arquiteturas de Integração Web SGBD: Um Estudo do Ponto de Vista de Sistemas de BANCO DE DADOS,** Iremar Nunes de Lima, Sergio Lifschitz
- **[1997] Interoperabilidade em um Sistema de BANCOS DE DADOS Heterogêneos usando padrão CORBA,** Elvira Maria Antunes Uchôa, Rubens Nascimento Melo, Sergio Lifschitz

Figura 4. Result of use of wildcards

After selecting a professor/researcher, it is possible to read their biography, access their contact page, their Lattes Curriculum, and the productions and lectures, if any. When selecting any production categories, the items containing the searched word appear first, and then the articles released in the years 2018, 2019, and 2020 appear ordered in reverse

chronological order.

In addition, the system enables to search for a professor/researcher name directly, in which the user can enter the full name or part of it. We were able to develop a way to understand when the user is searching for a keyword or the professor/researcher's name, bringing more semantics to the project as presented in figure 5. According to the triple predicate, the SPARQL query returns if the keyword matches with a person name or a publication title.

Figura 5. Result of desambiguation

5. Conclusions

The publication of this tool on the Web represents an efficiency gain in identifying professors/researchers from the PUC-Rio community. But Quem@PUC is yet not able to deal with the ambiguity of words. The tool focuses on the terms involved in the searches (syntax) and not on their meanings (semantics). To overcome this limitation, it will be necessary to better incorporate semantic elements in the search engine to understand the users' intentions behind their searches. Among the possibilities to carry out this incorporation, there is a controlled list of keywords, thesaurus, domain ontologies, and even knowledge graphs that are being studied at the present moment.

Referências

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern information retrieval. In *Modern Information Retrieval*, pages 1–2. Pearson, Addison-Wesley.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Ziegler, P. and Dittrich, K. R. (2007). Data integration – problems, approaches, and perspectives. In Krogstie, J., Opdahl, A. L., and Brinkkemper, S., editors, *Conceptual Modelling in Information Systems Engineering*, pages 39–58. Springer.