Sistema de Recuperação de Informações Busc@NIMA

Aluno: Mariana Duarte de Araujo Salgueiro
Orientador: Sérgio Lifschitz

Introdução

O NIMA (Núcleo Interdisciplinar de Meio Ambiente) [1] é uma unidade da PUC-Rio que tem por objetivo ser o local de discussões interdisciplinares sobre as questões socioambientais. O NIMA desenvolve projetos e pesquisas sobre assuntos relacionados ao meio ambiente dentro da PUC-Rio, ou em cooperação com outras instituições de ensino e pesquisa nacionais e internacionais. Este projeto teve por objetivo desenvolver um Sistema de Recuperação de Informações (*information retrieval*) [2] que identifique projetos e trabalhos de pesquisa e/ou desenvolvimento realizados na PUC-Rio envolvendo professores, funcionários e alunos. A PUC-Rio, por meio deste sistema a ser abrigado no website do NIMA, poderá assim divulgar, para a sociedade em geral, ou pesquisadores em particular, as suas atividades na área do meio ambiente, e as competências existentes em seus laboratórios e departamentos.

A ferramenta, denominada Busc@NIMA, funciona como um site de buscas tradicional, porém contemplando apenas as produções encontradas nos CV Lattes [3] dos membros da comunidade PUC-Rio. É possível obter informações sobre professores, tais como suas páginas pessoais em websites departamentais, e também dados sobre as disciplinas oferecidas na PUC-Rio e os professores envolvidos.

Através do SGBD (Sistema Gerenciador de Bancos de Dados) AllegroGraph [4], do *microframework* Flask [5], e da linguagem Python [6], foi possível construir uma ferramenta para acesso web capaz de disponibilizar informações relacionadas com uma determinada palavra-chave buscada. A ferramenta não exige autenticação de seus usuários e pode ser acessada livremente em qualquer lugar do mundo, bastando ter acesso à internet.

Figura 1 - CV Lattes, websites de professores e disciplinas PUC-Rio compõem a base de dados da ferramenta Busc@NIMA

Web Semântica

A estrutura da linguagem HTML (HyperText Markup Language) [7], utilizada hoje nas páginas web, é rígida, e enfatiza a estrutura e forma de exibição de documentos para navegadores (*browsers*). A adição de novos comandos de marcação, para uma melhor descrição de uma página web, geraria a necessidade da redefinição do DTD (Definição de Tipo de Documento) [8] da linguagem e, consequentemente, uma atualização dos navegadores web para que interpretassem estes novos comandos. A linguagem XML (Extensible Markup Language) [9] foi criada a partir desta limitação da HTML e da necessidade de uma linguagem que pudesse descrever também o conteúdo semântico e os significados contextuais de documentos na web. De fato, a HTML tem por objetivo controlar

a forma com que os dados serão exibidos e a XML se concentra mais na descrição dos dados que o documento contém [10].

O RDF (*Resource Description Framework*) é um modelo padrão [11] que permite que dados sejam compartilhados e reutilizados com o apoio de ontologias [12] e que deve vir a ser implementado na confecção de páginas da Web Semântica [13]. Este modelo possui recursos que facilitam a integração e a interoperabilidade de dados, e estabelece um padrão de metadados para ser embutido na codificação XML. A ideia do RDF é a descrição dos dados e dos metadados por meio de um esquema de **triplas** <S, P, O> (**sujeito** -> **predicado** -> **objeto**). Os **sujeitos** podem ser URIs (*Uniform Resource Identifier*), usados para identificar ou denominar um recurso na internet, ou nós em branco, estruturas em RDF em que uma classe tem relacionamento com diversas outras classes [14]; **objetos** podem ser URIs, nós em branco ou literais; e **predicados** são elementos de ligação (geralmente declarados em uma ontologia) que relacionam estes recursos, como é feito em um grafo direcionado, conectado através de suas arestas.

As ontologias definem vocabulários controlados que identificam um conjunto de conceitos de forma única para que não haja ambiguidade na sua interpretação. Portanto, uma ontologia também pode ser vista como uma categoria de modelo de dados que define formalmente os relacionamentos entre conceitos. As ontologias visam estabelecer uma relação organizada e padronizada entre termos, favorecendo a contextualização dos dados, extraídos de diversas fontes, e facilitando o processo de interpretação dos mesmos.

Por conta disso, torna-se cada vez mais interessante que os buscadores incorporem elementos semânticos em seus mecanismos com o intuito de melhor compreender as intenções dos usuários por trás de suas buscas [15] e a ferramenta Busc@NIMA, aqui desenvolvida, está projetada seguindo estes requisitos.

Conversão dos CV Lattes

A conversão dos CV Lattes dos professores da PUC-Rio se deu através de um *parser* que transforma o XML do currículo em RDF, usando XSLT [16]. Entretanto, este *parser* inicial possui um *script* de conversão XML para RDF feito somente para mapeamento de produções científicas (artigos em conferências e revistas, livros, teses, capítulos de livro etc.) e para a ferramenta Busc@NIMA seriam interessantes disponibilizarmos outras informações relevantes contidas nos CV Lattes, como projetos de pesquisa e desenvolvimento, orientações de trabalhos de graduação e pós-graduação, dentre outros. As Figuras 2 e 3, a seguir, ilustram esta conversão de trechos de um CV Lattes em um esquema RDF.

```
<ARTIGO-PUBLICADO SEQUENCIA-PRODUCAO="78" ORDEM-IMPORTANCIA="">
    <DADOS-BASICOS-DO-ARTIGO NATUREZA="COMPLETO" TITULO-DO-ARTIGO="A terceirização sob o prisma do trabalho e do desenvolvimento social."
    ANO-DO-ARTIGO="2014" PAIS-DE-PUBLICACAO="" IDIOMA="Português" MEIO-DE-DIVULGACAO="VARIOS" HOME-PAGE-DO-TRABALHO="" FLAG-RELEVANCIA="NAO" DOI=""
    TITULO-DO-ARTIGO-INGLES="" FLAG-DIVULGACAO-CIENTIFICA="NAO"/>
    <DETALHAMENTO-DO-ARTIGO TITULO-DO-PERIODICO-OU-REVISTA="Revista do Tribunal Superior do Trabalho" ISSN="01037978" VOLUME="80" FASCICULO="" SERIE="3"
    PAGINA-INICIAL="257" PAGINA-FINAL="267" LOCAL-DE-PUBLICACAO=""/>
    <AUTORES NOME-COMPLETO-DO-AUTOR="Sayonara Grillo Coutinho Leonardo da Silva" NOME-PARA-CITACAO="SILVA, S. G. C. L." ORDEM-DE-AUTORIA="1"
    NRO-ID-CNPQ="0059048013298492"/>
    <AUTORES NOME-COMPLETO-DO-AUTOR="Ana Luisa de Souza Correia de Melo Palmisciano" NOME-PARA-CITACAO="PALMISCIANO, A.L.S.C.M." ORDEM-DE-AUTORIA="2"
    NRO-ID-CNPQ="0080400691590938"/>
</ARTIGO-PUBLICADO>
```

Figura 2 - Exemplo de produção (artigo) em XML

```
<rdf:Description rdf:about="#P78">
  <rdf:type rdf:resource="http://purl.org/ontology/bibo/Article"/>
  <dc:title>A terceirização sob o prisma do trabalho e do desenvolvimento social.</dc:title>
  <dcterms:issued>2014</dcterms:issued>
  <dc:language>Português</dc:language>
  <dcterms:isPartOf>
    <rdf:Description>
      <rdf:type rdf:resource="http://purl.org/ontology/bibo/Journal"/>
      <bibo:issn>01037978</bibo:issn>
      <dc:title>Revista do Tribunal Superior do Trabalho</dc:title>
    </rdf:Description>
  </dcterms:isPartOf>
  <bibo:pageStart>257</bibo:pageStart>
  <bibo:pageEnd>267</bibo:pageEnd>
  <bibo:volume>80</bibo:volume>
  <dc:creator rdf:resource="#author-idm46124418089568"/>
  <dc:creator rdf:resource="#author-idm46124418088352"/>
  <bibo:authorList rdf:parseType="Collection">
    <rdf:Description rdf:about="#author-idm46124418089568"/>
    <rdf:Description rdf:about="#author-idm46124418088352"/>
  </bibo:authorList>
  <dcterms:isReferencedBy rdf:resource=""/>
</rdf:Description>
```

Figura 3 - Conversão do artigo da Figura 2 em RDF através do script XLST

Banco de Dados e Ferramenta de ETL

Um dos desafios propostos neste projeto de pesquisa foi, exatamente, sair do escopo dos tradicionais sistemas de bancos de dados relacionais para os sistemas de bancos de dados conhecidos como NoSQL (*Not-only* SQL ou não-relacionais) [17]. Isto foi motivado também pela variedade de fontes de dados que compõem a base da nossa ferramenta de busca, não sendo possível estabelecer uma modelagem esquemática a priori, e pela oportunidade de explorar novas tecnologias não estudadas em cursos regulares de graduação em computação. Uma modelagem de dados sem esquema rígido permite que os dados tenham estruturas diversas, sem ou com poucas regras ou normas. Essa flexibilidade é muito importante para o projeto Busc@NIMA devido ao fato de que diversas fontes de dados são consideradas.

O SGBD não-relacional AllegroGraph é uma *triplestore* [18] projetada para armazenar triplas RDF e manipulação, assim como visualização, como estruturas de grafo. Trata-se, na realidade, de um sistema de banco de dados considerado *polystore* ou multi-modelo [19] (documentos em JSON, JSON-LD [20] e grafos em RDF e OWL [21]) que atende aos padrões W3C [22] para a Web Semântica. AllegroGraph oferece suporte à linguagem SPARQL [23], que é uma linguagem de consulta padrão para *linked data* [24] e, também, há suporte nativo à indexação de dados textuais (*free text search*) permitindo mapear rapidamente palavras e frases para as triplas do banco de dados.

A API AllegroGraph Python API [25] foi utilizada neste projeto por fornecer acesso eficiente ao SGBD AllegroGraph para aplicações escritas na linguagem Python. Essa API fornece métodos para criar, consultar e manter dados RDF, e para gerenciar as triplas armazenadas.

ETL (*Extract, Transform and Load)* é um processo de integração de dados em três etapas usado tipicamente para combinar dados de diversas fontes [26]. A **extração** consiste na coleta e importação de dados de diversos sistemas. Já a **transformação** consiste no mapeamento dos dados extraídos conforme regras de negócio definidas. Por fim, a **carga** consiste na última etapa da integração, com o armazenamento dos dados no repositório destino.

O processo de conversão das informações dos professores e das disciplinas oferecidas na PUC-Rio foi feito através da ferramenta de carga de ETL *Linked Pipes* [27]. É uma ferramenta leve, especializada na conversão e carga de *Linked Open Data* [28]. Possui vários

componentes de conversão (CSV para RDF, XML para RDF etc.) além dos componentes para manipulação de dados em formato RDF usando SPARQL. Esse tipo de ferramenta já possui funções específicas para a criação de processos automatizados de conversão e carga de dados. Não é necessário conhecimento de linguagens de programação de baixo nível. São usados os metadados e logs de execução dos processos gerados e mantidos no catálogo da própria ferramenta, bem como podem ser instalados em servidores separados dos servidores da aplicação, permitindo a conexão à repositórios remotos (como fonte ou destino dos dados).

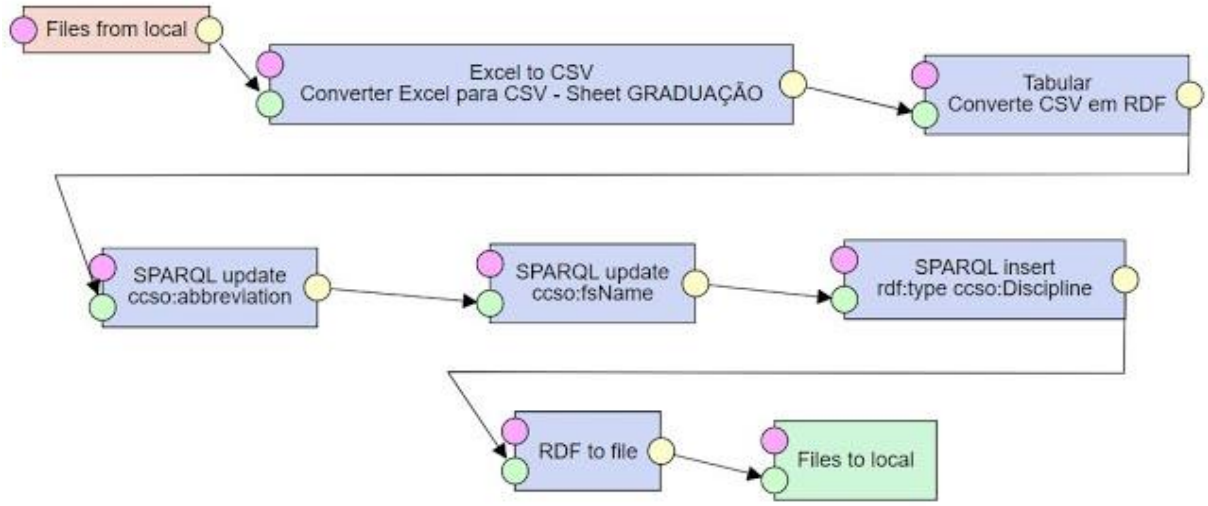


Figura 4 - Exemplo de uso da ferramenta Linked Pipes para conversão de arquivo Excel de disciplinas da PUC-Rio em RDF

Foi criado um *pipeline* na ferramenta *Linked Pipes* para conversão em RDF do arquivo Excel de disciplinas representado na Figura 4. No caso, os passos são os seguintes:

i. **Excel to CSV** é um conversor de arquivos de planilhas eletrônica em formato MS Excel para CSV (*comma separated values*) que gera um arquivo CSV para cada planilha do arquivo de entrada.
ii. **Tabular** é um conversor de CSV para RDF que segue o padrão da W3C.
iii. Dois componentes **SPARQL update** foram utilizados para mapear as colunas Código e Nome das disciplinas de graduação nas propriedades *ccso:abbreviation* e *ccso:fsName* da ontologia *Curriculum Course Syllabus Ontology* (CCSO) [29].
iv. Um terceiro componente *SPARQL update* e o comando SPARQL (Figura 5), foram utilizados para relacionar as instâncias de Disciplinas com a classe *Discipline*:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ccso: <https://w3id.org/ccso/ccso#>
INSERT {?s rdf:type ccso:Discipline .}
WHERE  {?s ?p ?o .
      FILTER (?p = <https://w3id.org/ccso/ccso#abbreviation>) .
}
```

Figura 5 - Comando SPARQL para relacionar as instâncias de Disciplinas com a classe *Discipline*.

O arquivo final gerado pelo *pipeline* foi copiado para o servidor do projeto.

Atividades Realizadas

Foi necessário, inicialmente, fazer diversas pesquisas sobre RDF para entender seu conceito e sua usabilidade. Os conceitos de OWL, Open Linked Data, Web Semântica, XML, ontologias, bancos de dados NoSQL e SPARQL, não cobertos por disciplinas regulares de graduação em computação, também precisaram ser estudados.

O *parser* de conversão de CV Lattes em XML para RDF foi testado para uma base de dados de CV Lattes de novembro de 2019 com a motivação de uma prova de conceito.

```
echo "Iniciando o script:"
cd lattes-professores-xml/

echo "Convertendo os xmls em rdfs..."
for f in *.xml
do
    newFile=${f/${f: -4}/}

    # need to install libxml2, libxslt
    xsltproc --stringparam ID $newFile ../lattes_professores.xsl $f > $newFile".rdf"
done

echo "Movendo rdfs para diretorio separado ..."
find . -name "*.rdf" -type f | head -546 | while read arq; do mv "$arq" ../lattes-professores-rdf/; done;
mv *.rdf ../lattes-professores2-rdf/
```

Figura 5 - Script bash para conversão XML-RDF

O SGBD AllegroGraph precisou ser estudado para se entender como seria possível conectá-lo à aplicação feita na linguagem Python e, também, para entender como seriam feitas as consultas para recuperar as informações desejadas. Sua instalação no servidor do projeto foi feita sem dificuldades seguindo os comandos da documentação [30] disponível. A interação pode ser feita através de duas interfaces: (i) via *browser*, chamada de *WebView* [31], onde o usuário pode explorar, consultar e gerenciar repositórios de maneira mais clara, ou (ii) via linha de comando. A versão do banco de dados utilizada é gratuita. Entretanto, existe uma limitação de no máximo 5 milhões de triplas por repositório criado, o que se mostrou insuficiente. Depois da conversão de todos os CVs Lattes de membros da comunidade PUC-Rio, o volume de triplas ficou muito grande e foi necessário criar dois repositórios separados. Estes repositórios respeitam o domínio da informação (separação por fonte de dados) e a limitação do tamanho especificada na versão utilizada.

Figura 6 - AllegroGraph WebView

Para que a busca pela palavra-chave indicada na página principal da aplicação retorne um resultado que envolva todos os repositórios considerados, foi necessário implementar *threads* [32] com a biblioteca *Threading* [33] do Python. A consulta principal em SPARQL roda simultaneamente nos repositórios e os resultados são unidos através de um dicionário.

Como alternativa para explorar os dados e, também, para montar as consultas em SPARQL, a ferramenta AllegroGraph oferece o Gruff [34], uma ferramenta desktop interativa para navegar, consultar e editar triplas. A representação dos dados é feita em forma de grafos para melhor entendimento dos dados armazenados no banco (Figura 7).

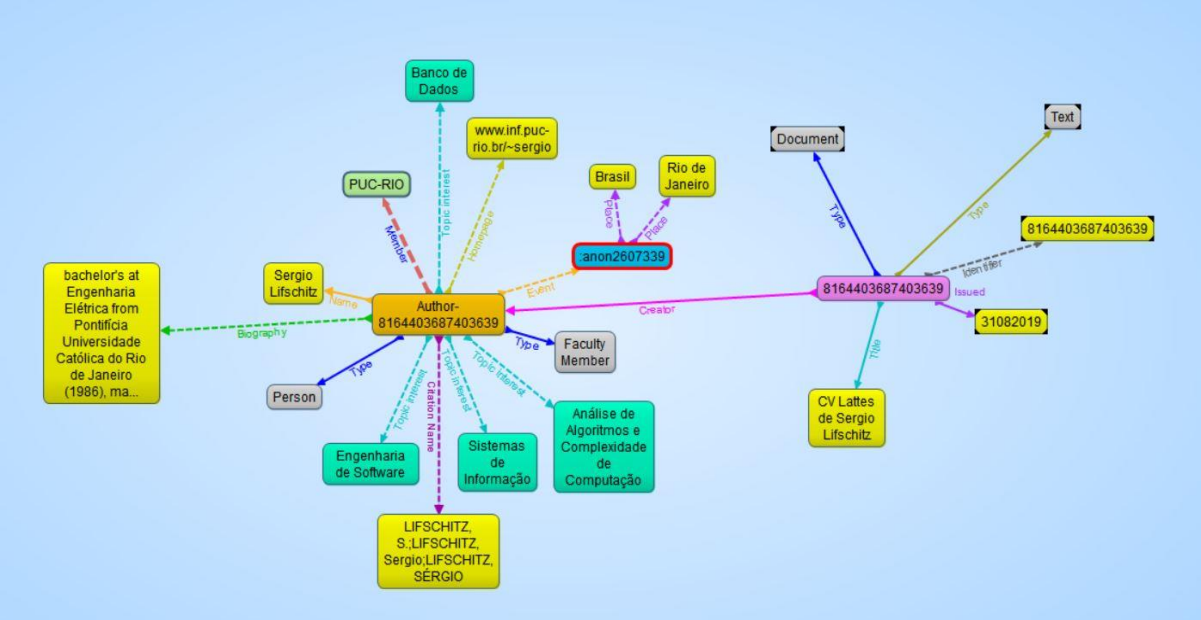


Figura 7 - Visualização em forma de grafo de um sujeito e seus predicados e objetos

Foi feita uma grande pesquisa e estudo de artigos envolvendo o tópico de Sistemas de Recuperação de Informação, de como estes sistemas se relacionam com a Web Semântica e de como os bancos de dados conhecidos como *triplestores* são capacitados para armazenarem representações semânticas de dados. Os CV Lattes que compõem o banco de dados do projeto, serão atualizados periodicamente, em função da liberação de novas extrações realizadas pela CCPA (Coordenação Central de Planejamento Acadêmico) da PUC-Rio. No caso da CCPA, é possível obter um conjunto de dados de CV Lattes mais facilmente do que usuários comuns, que precisam solicitar um de cada vez, controlado pelo reCaptcha [35].

Infraestrutura

A construção da ferramenta foi feita com uso da linguagem de programação Python em conjunto com o *microframework* Flask [36], que permite uma rápida prototipação de aplicações web do tipo CRUD (Create. Read, Update, Delete) [37]. O Flask é baseado nos pacotes Werkzeug [38], que oferece suporte no gerenciamento de requisições ao website, e na geração das páginas HTML, através do Jinja2 [39]. Flask oferece, também, flexibilidade quanto à persistência de dados e trabalha com componentes "plugáveis" (extensões).

A construção de um projeto em Flask exige uma determinada organização de pastas e módulos. Apesar de complexa, a estrutura permite organizar os módulos do projeto de maneira lógica, ilustrada, por exemplo, na Figura 8.

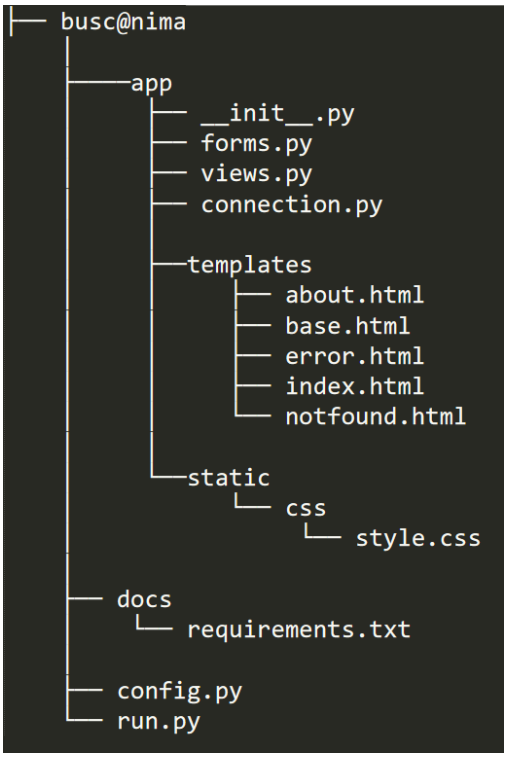


Figura 8 - Arquitetura Flask

A tabela a seguir explica o conteúdo de cada arquivo/pasta presente na estrutura:

run.py	Arquivo que, após carregar o objeto app do projeto (trazendo consigo todas as configurações da aplicação) executa a aplicação em um servidor local. Serve apenas para desenvolvimento.
/docs/requirements.txt	Arquivo que contém todos os pacotes do Python necessários para que a aplicação funcione.
config.py	Arquivo que contém todas as variáveis de configuração da aplicação.
/app	Pasta que contém o material (código Python, Javascript, HTML, CSS etc) relacionados à aplicação.
/app/__init__.py	Arquivo que inicializa um módulo (neste caso, usado para inicializar a aplicação geral).
/app/views.py	Arquivo onde são definidas as rotas (URLs) do projeto e o que acontece (em termos de código) quando elas são acessadas pelo usuário.
/app/static/	Pasta que contém todo o material público da aplicação, como código Javascript, CSS.
/app/templates/	Pasta que contém todos os arquivos HTML de cada página do site.

/app/forms.py	Definição do formulário de dados quando é necessário trabalhar com o envio deles para o backend por um navegador.
/app/connection.py	Arquivo de conexão com o banco de dados AllegroGraph.

O sistema todo está dividido em dois servidores: um contendo a ferramenta de ETL, onde são feitas todas as conversões necessárias para RDF (ETL Server), e o outro um servidor de aplicação (Web Server + DB Server), que recebe estes arquivos convertidos e gera toda a aplicação em si. No servidor de aplicação é feita a carga no banco de dados AllegroGraph e a hospedagem do site gerenciado com a ferramenta Apache [40].

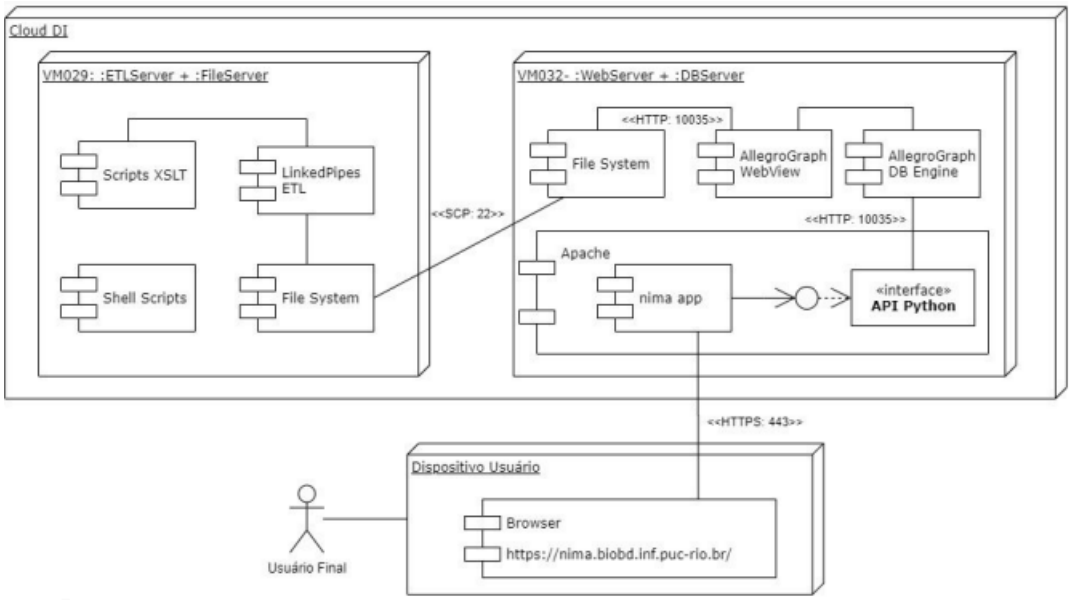


Figura 9 - Diagrama de Implantação

Ferramenta de Busca

A aplicação web desenvolvida neste projeto está disponível para testes em https://nima.biobd.inf.puc-rio.br/ e já conta com usuários selecionados que fazem testes visando ajustes de interface e conteúdo dos resultados. Nesta aplicação o usuário pode realizar uma busca sintática por palavra-chave com filtros por tipo de recurso de interesse (artigos/livros/capítulos ou teses).

Por ora, o resultado das buscas inclui a lista de membros da comunidade PUC-Rio relacionados com o termo buscado, ordenados por um critério de relevância temporariamente definido em função da quantidade de documentos associados com as pessoas na base de CV Lattes e disciplinas da PUC-Rio. Em caso de eventuais problemas encontrados por parte dos usuários, a aplicação está *logada* em um arquivo sequencial indicando os erros, de forma que as devidas correções podem ser feitas rapidamente.

Figura 10 - Interface de busca por palavra-chave

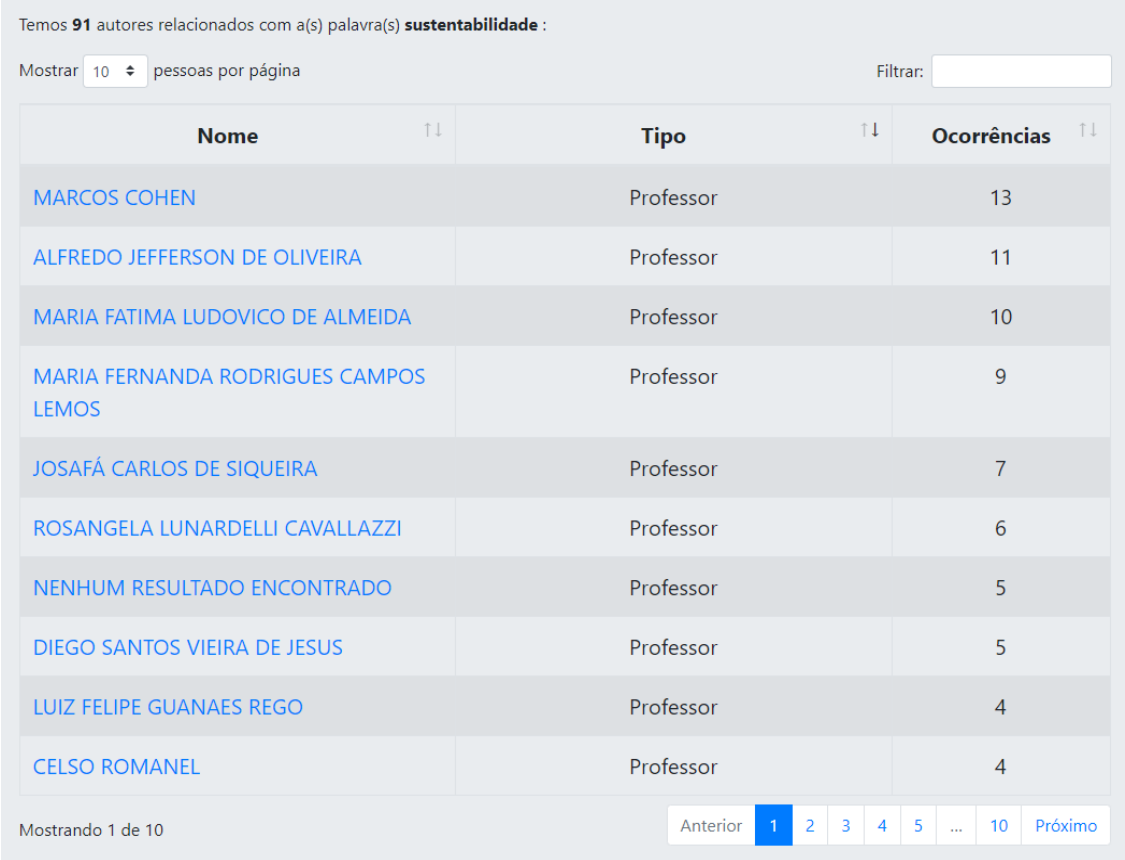
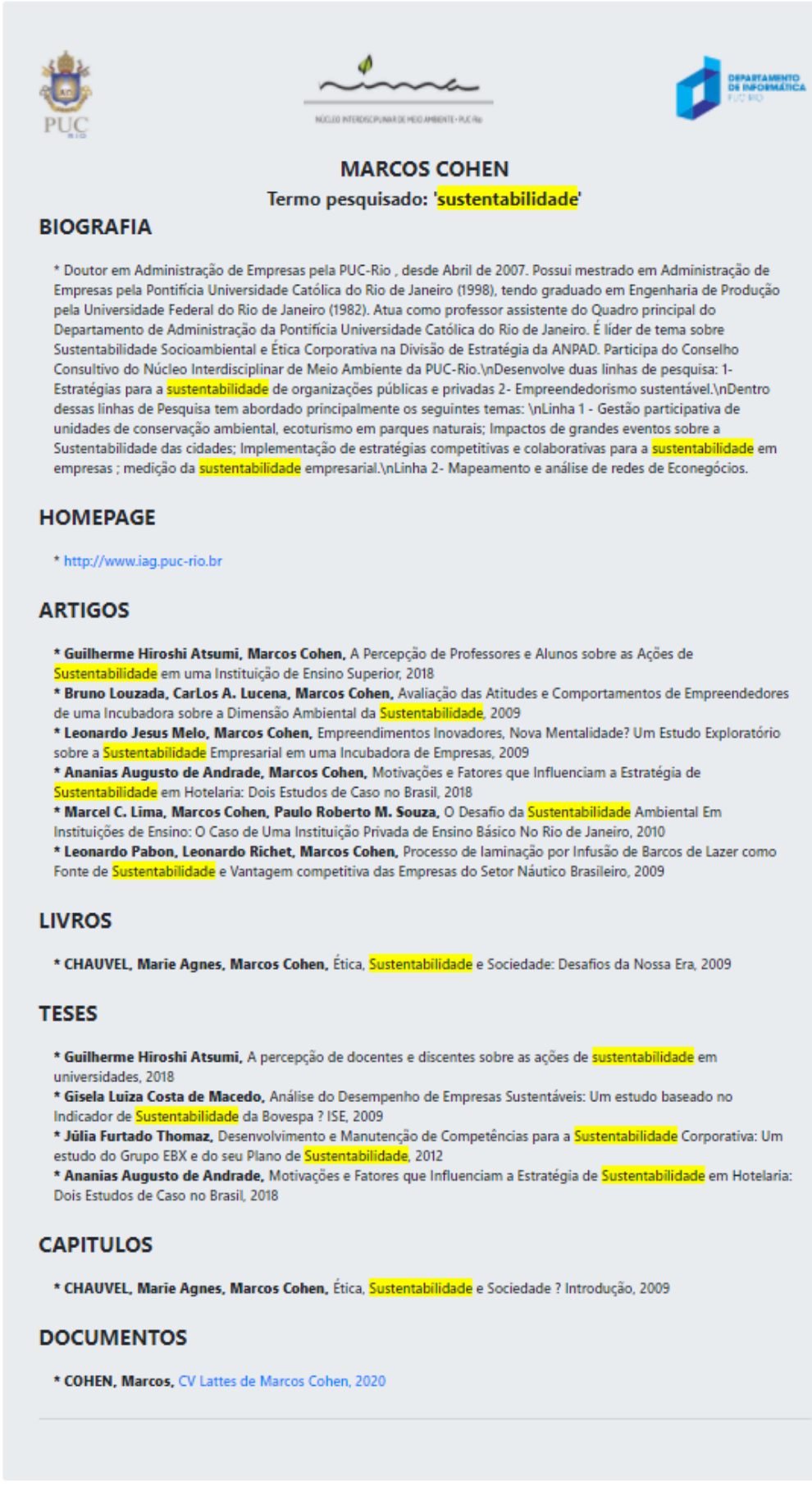


Figura 11 - Resultado da busca



MARCOS COHEN

Termo pesquisado: 'sustentabilidade'

BIOGRAFIA

* Doutor em Administração de Empresas pela PUC-Rio , desde Abril de 2007. Possui mestrado em Administração de Empresas pela Pontifícia Universidade Católica do Rio de Janeiro (1998), tendo graduado em Engenharia de Produção pela Universidade Federal do Rio de Janeiro (1982). Atua como professor assistente do Quadro principal do Departamento de Administração da Pontifícia Universidade Católica do Rio de Janeiro. É líder de tema sobre Sustentabilidade Socioambiental e Ética Corporativa na Divisão de Estratégia da ANPAD. Participa do Conselho Consultivo do Núcleo Interdisciplinar de Meio Ambiente da PUC-Rio.\nDesenvolve duas linhas de pesquisa: 1- Estratégias para a sustentabilidade de organizações públicas e privadas 2- Empreendedorismo sustentável.\nDentro dessas linhas de Pesquisa tem abordado principalmente os seguintes temas: \nLinha 1 - Gestão participativa de unidades de conservação ambiental, ecoturismo em parques naturais; Impactos de grandes eventos sobre a Sustentabilidade das cidades; Implementação de estratégias competitivas e colaborativas para a sustentabilidade em empresas ; medição da sustentabilidade empresarial.\nLinha 2- Mapeamento e análise de redes de Econegócios.

HOMEPAGE

* http://www.iag.puc-rio.br

ARTIGOS

* **Guilherme Hiroshi Atsumi, Marcos Cohen,** A Percepção de Professores e Alunos sobre as Ações de Sustentabilidade em uma Instituição de Ensino Superior, 2018
* **Bruno Louzada, CarLos A. Lucena, Marcos Cohen,** Avaliação das Atitudes e Comportamentos de Empreendedores de uma Incubadora sobre a Dimensão Ambiental da Sustentabilidade, 2009
* **Leonardo Jesus Melo, Marcos Cohen,** Empreendimentos Inovadores, Nova Mentalidade? Um Estudo Exploratório sobre a Sustentabilidade Empresarial em uma Incubadora de Empresas, 2009
* **Ananias Augusto de Andrade, Marcos Cohen,** Motivações e Fatores que Influenciam a Estratégia de Sustentabilidade em Hotelaria: Dois Estudos de Caso no Brasil, 2018
* **Marcel C. Lima, Marcos Cohen, Paulo Roberto M. Souza,** O Desafio da Sustentabilidade Ambiental Em Instituições de Ensino: O Caso de Uma Instituição Privada de Ensino Básico No Rio de Janeiro, 2010
* **Leonardo Pabon, Leonardo Richet, Marcos Cohen,** Processo de laminação por Infusão de Barcos de Lazer como Fonte de Sustentabilidade e Vantagem competitiva das Empresas do Setor Náutico Brasileiro, 2009

LIVROS

* **CHAUVEL, Marie Agnes, Marcos Cohen,** Ética, Sustentabilidade e Sociedade: Desafios da Nossa Era, 2009

TESES

* **Guilherme Hiroshi Atsumi,** A percepção de docentes e discentes sobre as ações de sustentabilidade em universidades, 2018
* **Gisela Luiza Costa de Macedo,** Análise do Desempenho de Empresas Sustentáveis: Um estudo baseado no Indicador de Sustentabilidade da Bovespa ? ISE, 2009
* **Júlia Furtado Thomaz,** Desenvolvimento e Manutenção de Competências para a Sustentabilidade Corporativa: Um estudo do Grupo EBX e do seu Plano de Sustentabilidade, 2012
* **Ananias Augusto de Andrade,** Motivações e Fatores que Influenciam a Estratégia de Sustentabilidade em Hotelaria: Dois Estudos de Caso no Brasil, 2018

CAPITULOS

* **CHAUVEL, Marie Agnes, Marcos Cohen,** Ética, Sustentabilidade e Sociedade ? Introdução, 2009

DOCUMENTOS

* **COHEN, Marcos,** CV Lattes de Marcos Cohen, 2020

Figura 12 - Detalhamento do pesquisador

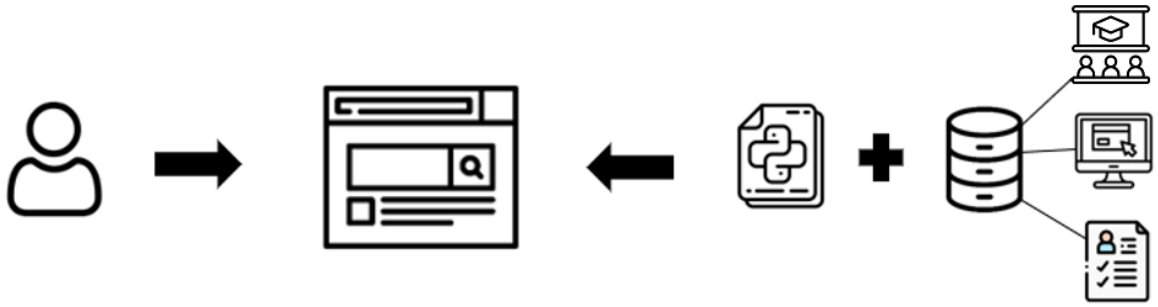


Figura 13 - Arquitetura Busc@NIMA

Desafios Encontrados

1. Qualidade das fontes de dados:
    a. O processo de *triplificação* pode criar sujeitos, nas triplas, que não possuem vínculo com a PUC-Rio, como é o caso de coautores de artigos membros de outras universidades. Foi preciso filtrar para que não se obtenha um número enorme de triplas referentes a pessoas que nem deveriam aparecer na aplicação.
    b. O *matching* entre os professores listados no arquivo de disciplinas e seus CV Lattes precisa de HITL [41] para que possamos identificar e atribuir informações corretas às pessoas certas.
    c. Inicialmente seriam considerados os CV Lattes tanto de professores como de alunos da PUC-Rio, mas o conjunto XML é muito incompleto por ser não ser comum que alunos informem seus identificadores. Optamos então, ao menos temporariamente, por remover o repositório de alunos do banco de dados.

2. Modelagem conceitual
   Entender as fontes, entender o que cada entidade envolvida representa para que se fosse possível fazer um mapeamento, entender a relação dos campos dos arquivos envolvidos para poder adicionar semântica exige estudo aprofundado. A ideia que embasa o projeto é não fazer uma conversão só de formato, já que se pode ter outros usos para esses dados, e sim manter a semântica junto com dado exigiu grande esforço no desenvolvimento do projeto.

3. Tecnologias (XSLT)
   Precisamos extrair mais informações dos CV Lattes já que o *parser* até hoje utilizado só extrai produções envolvendo artigos em conferências e revistas. Aprender a mexer com XSLT é um grande desafio.

4. Recursos computacionais
   A VM (máquina virtual) disponibilizada para o projeto possuía 8GB de memória RAM e com a introdução de novos repositórios de dados, a memória foi consumida quase em sua totalidade. Como existem milhões de triplas RDF tivemos que aumentar a memória de 8 para 16GB.

5. Versão do SGBD AllegroGraph
   A versão do SGBD utilizada desde o início do projeto foi a 6.4.2 e o planejamento era atualizá-la para a 7.0.2. A atualização traria algumas funcionalidades que poderiam ser utilizadas, como as Federações – que servem para substituir o uso de *threads*, já que fazem a união entre todos os repositórios, permitindo que quaisquer consultas sejam executadas de uma só vez. Entretanto, dentro do período de execução deste projeto isto não pôde ser feito devido à necessidade de recursos computacionais ainda não disponíveis.

Conclusões

A ferramenta Busc@NIMA, então, é um protótipo de Sistema de Recuperação de Informações que identifica projetos e trabalhos de pesquisa e/ou desenvolvimento de professores envolvidos com a temática de meio ambiente na PUC-Rio.

O objetivo inicial do projeto era desenvolver a ferramenta contemplando os CV Lattes, as informações de professores (tais como suas páginas pessoais em websites departamentais) e os dados sobre as disciplinas oferecidas na PUC-Rio e os professores envolvidos. Destas três fontes principais, não foi possível incluir as disciplinas oferecidas devido a necessidade de atualização do SGBD AllegroGraph e do desenvolvimento de programas de apoio para que fosse possível fazer o *match* entre professores já existentes na base de dados e os que lecionam as disciplinas levantadas. Esta questão será contemplada nos próximos passos do projeto, assim como o desenvolvimento de um XSLT que transforme não somente produções do CV Lattes em RDF, mas também, projetos de pesquisa e desenvolvimento, orientações de trabalhos de graduação e pós-graduação, dentre outros.

O aprendizado ao longo do desenvolvimento do projeto foi significativo já que explorou diversas áreas desconhecidas previamente por mim. Agradeço à orientação do professor Sérgio Lifschitz, a disponibilização da ferramenta XLST pelos professores Jefferson Santos e Edward Hermann Haeusler, e as alunas Veronica dos Santos, doutoranda do DI, e Andrea Mourelo Rodriguez, aluna da Ecole CentraleSupelec (Paris, França), que participaram do desenvolvimento da Busc@NIMA. Cabe também um agradecimento especial ao funcionário da CCPA PUC-Rio e ex-aluno do DI, Tomás Guisasola, que prepara e disponibiliza as planilhas com CVs Lattes atualizados envolvendo professores da PUC-Rio. Sem sua ajuda teríamos que obter os CVs Lattes toda vez passando pelo reCaptcha, o que tornaria o processo muito demorado.

Referências

1 – NIMA – Núcleo Interdisciplinar de Meio Ambiente da PUC-Rio. URL: <http://www.nima.puc-rio.br>. Acesso em: 3 set. 2020.

2 – CARDOSO, Olinda Nogueira Paes. Recuperação de Informação. **INFOCOMP Journal of Computer Science**, v. 2, n. 1, p. 33-38, 2004.

3 – Plataforma Lattes. Lattes.cnpq.br. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 3 set. 2020.

4 – AllegroGraph. Allegrograph.com. Disponível em: <https://allegrograph.com/>. Acesso em: 3 set. 2020.

5 – Welcome to Flask — Flask Documentation (1.1.x). Flask.palletsprojects.com. Disponível em: <https://flask.palletsprojects.com/en/1.1.x/>. Acesso em: 3 set. 2020.

6 – Welcome to Python.org. Python.org. Disponível em: <https://www.python.org/>. Acesso em: 3 set. 2020.

7 – HTML Standard. Html.spec.whatwg.org. Disponível em: <https://html.spec.whatwg.org/multipage/>. Acesso em: 3 set. 2020.

8 – Definição de Tipo de Documento. Pt.wikipedia.org. Disponível em: <https://pt.wikipedia.org/wiki/Defini%C3%A7%C3%A3o_de_Tipo_de_Documento>. Acesso em: 3 set. 2020.

9 – Extensible Markup Language (XML). W3.org. Disponível em: <https://www.w3.org/XML/>. Acesso em: 3 set. 2020.

10 – SOUZA, Renato Rocha; ALVARENGA, Lídia. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, v. 33, n. 1, 2004.

11 – RDF - Semantic Web Standards. W3.org. Disponível em: <https://www.w3.org/RDF/>. Acesso em: 4 jul. 2020.

12 – DE FREITAS, Frederico Luiz Gonçalves. Ontologias e a web semântica. **Jornada de Mini-Cursos em Inteligência Artificial, SBC**, v. 8, 2003.

13 – The original proposal of the WWW, HTMLized. W3.org. Disponível em: <https://www.w3.org/History/1989/proposal.html>. Acesso em: 3 set. 2020.

14 – DE LIMA, Júnio César; DE CARVALHO, Cedric L. **Resource description framework (rdf)**. Technical report, Universidade Federal de Goiás, 2005.

15 – ROZSA, Vitor; GODOY VIERA, Angel Freddy; DUTRA, Moisés. Aplicação de Tecnologias da Web Semântica em Motores de Busca na Internet. **Investigación bibliotecológica**, v. 33, n. 78, p. 165-191, 2019.

16 – Como: usar XSLT Transformations com arquivos de intercâmbio de dados XML do Project. Docs.microsoft.com. Disponível em: <https://docs.microsoft.com/pt-br/office-project/xml-data-interchange/how-to-use-xslt-transformations-with-project-xml-data-interchange-files?view=project-client-2016#:~:text=XSLT%20%C3%A9%20usado%20para%20transformar,ou%20em%20o>. Acesso em: 5 jul. 2020.

17 – DE DIANA, Mauricio; GEROSA, Marco Aurélio. Nosql na web 2.0: Um estudo comparativo de bancos não-relacionais para armazenamento de dados na web 2.0. In: **IX Workshop de Teses e Dissertações em Banco de dados**. 2010.

18 – What is RDF Triplestore? - Ontotext. Ontotext. Disponível em: <https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf-triplestore/>. Acesso em: 3 set. 2020.

19 – AQUINO, Amanda CV et al. Uma análise comparativa de sistemas de gerenciamento de bancos de dados NoSQL multimodelo. 2019.

20 – JSON-LD - JSON for Linking Data. Json-ld.org. Disponível em: <https://json-ld.org/>. Aesso em: 3 set. 2020.

21 – OWL - Semantic Web Standards. W3.org. Disponível em: <https://www.w3.org/OWL/>. Acesso em: 3 set. 2020.

22 – World Wide Web Consortium (W3C). W3.org. Disponível em: <https://www.w3.org/>. Acesso em: 3 set. 2020.

23 – SPARQL Query Language for RDF. W3.org. Disponível em: <https://www.w3.org/TR/rdf-sparql-query/>. Acesso em: 3 set. 2020.

24 – Data - W3C. W3.org. Disponível em: <https://www.w3.org/standards/semanticweb/data>. Acesso em: 3 set. 2020.

25 – Welcome to AllegroGraph Python client's documentation! — AllegroGraph Python client 100.0.5.dev0 documentation. Franz.com. Disponível em: <https://franz.com/agraph/support/documentation/6.4.0/python/index.html>. Acesso em: 1 jul. 2020.

26 – ETL: o que é e qual sua importância?. Sas.com. Disponível em: <https://www.sas.com/pt_br/insights/data-management/o-que-e-etl.html#:~:text=ETL%20%C3%A9%20um%20tipo%20de,combinar%20dados%20de>. Acesso em: 5 jul. 2020.

27 – ETL: Extract Transform Load for Linked Data. LikedPipes ETL. Disponível em: <https://etl.linkedpipes.com/>. Acesso em: 1 jul. 2020.

28 – Linked Open Data - W3C eGovernment Wiki. W3.org. Disponível em: <https://www.w3.org/egov/wiki/Linked_Open_Data>. Acesso em: 3 set. 2020.

29 – Curriculum Course Syllabus Ontology (CCSO). Vkreations.github.io. Disponível em: <https://vkreations.github.io/CCSO/>. Acesso em: 5 jul. 2020

30 – INCORPORATED, FRANZ. Server Installation | AllegroGraph 6.4.2. Franz.com. Disponível em: <https://franz.com/agraph/support/documentation/6.4.2/server-installation.html>. Acesso em: 10 ago. 2020.

31 – INCORPORATED, FRANZ. AllegroGraph WebView. Franz.com. Disponível em: <https://franz.com/agraph/support/documentation/current/agwebview.html>. Acesso em: 10 ago. 2020.

32 – KLEIMAN, Steve; SHAH, Devang; SMAALDERS, Bart. **Programming with threads**. Mountain View: Sun Soft Press, 1996.

33 – threading — Thread-based parallelism — Python 3.8.5 documentation. Docs.python.org. Disponível em: <https://docs.python.org/3/library/threading.html>. Acesso em: 10 ago. 2020.

34 – INCORPORATED, FRANZ. Gruff | AllegroGraph 7.0.0. Gruff.allegrograph.com. Disponível em: <http://gruff.allegrograph.com:10035/doc/gruff.html>. Acesso em: 10 ago. 2020.

35 – Welcome to Flask — Flask Documentation (1.1.x). Flask.palletsprojects.com. Disponível em: <https://flask.palletsprojects.com/en/1.1.x/>. Acesso em: 1 jul. 2020.

36 – Busca Textual - Curriculo Lattes. Buscatextual.cnpq.br. Disponível em: <http://buscatextual.cnpq.br/buscatextual/busca.do?metodo=apresentar>. Acesso em: 3 set. 2020.

37 – CRUD. Pt.wikipedia.org. Disponível em: <https://pt.wikipedia.org/wiki/CRUD>. Acesso em: 3 set. 2020.

38 – Werkzeug — Werkzeug Documentation (1.0.x). Werkzeug.palletsprojects.com. Disponível em: <https://werkzeug.palletsprojects.com/en/1.0.x/>. Acesso em: 1 jul. 2020.

39 – Jinja — Jinja Documentation (2.11.x). Jinja.palletsprojects.com. Disponível em: <https://jinja.palletsprojects.com/en/2.11.x/>. Acesso em: 1 jul. 2020.

40 – Welcome to The Apache Software Foundation!. Apache.org. Disponível em: <https://www.apache.org/>. Acesso em: 10 ago. 2020.

41 – Humans in the Loop: The Design of Interactive AI Systems. Disponível em: <https://hai.stanford.edu/blog/humans-loop-design-interactive-ai-systems>. Acesso em: 3 set. 2020.