
BioBD@PUC-Rio

Busc@NIMA

Especificação dos Requisitos de Sistema

Versão 1.3

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

Histórico da Revisão

Data	Versão	Descrição	Autor
01/05/2020	1.0	Versão Inicial	Veronica dos Santos
02/06/2020	1.1	Ajustes com os comentários do Sérgio e detalhamento de requisitos funcionais	Veronica dos Santos
10/06/2020	1.2	Finalização dos Casos de Uso de ETL	Veronica dos Santos
07/07/2020	1.3	Ajustes no UC1 depois do projeto de casos de teste e inclusão do informações de licenciamento	Veronica dos Santos

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

Índice Analítico

Introdução	5
Finalidade	5
Escopo	5
Definições, Acrônimos e Abreviações	6
Organização do documento	8
Requisitos	9
Funcionalidades	9
O sistema deve permitir que o usuário realize uma busca sintática simples por palavra-chave com filtro [RF1]	9
O sistema deve realizar o cálculo de indicador de relevância do pesquisador no tema Meio Ambiente [RF2]	9
O sistema deve permitir que o usuário realize uma busca lógica composta por palavra-chave com filtro [RF3]	10
O sistema deve permitir que o usuário realize uma busca semântica apoiada por Ontologia de Meio Ambiente [RF4]	10
O sistema deve realizar a conversão e carga de Currículo Lattes em base de dados triplicada [RF5]	10
O sistema deve realizar a conversão e carga de arquivo CSV com dados de cadastro de usuários em base de dados triplicada [RF6]	10
O sistema deve realizar a conversão e carga de planilha Excel com dados de disciplinas do SAU em base de dados triplicada [RF7]	11
O sistema deve realizar a triplicação das fontes de dados utilizando ontologias de domínio para atender aos padrões da Web Semântica [RF8]	11
O sistema deve realizar a correção ortográfica das palavras chaves digitadas na ferramenta de busca [RF9]	13
O sistema deve registrar as palavras chaves utilizadas pelos usuários nas ferramentas de busca [RF10]	13
O sistema deve fornecer ao administrador relatórios estatísticos de uso das palavras chaves de busca [RF11]	13
O sistema deve permitir ao pesquisador complementação de suas informações que não estão disponíveis publicamente assim como a correção de informações que se encontram na base de dados [RF12]	13
O sistema deve permitir ao administrador manter a lista de palavras chaves reservadas associadas ao tema Meio Ambiente [RF13]	14
Características e restrições do produto	14

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

Características e restrições do processo	15
Componentes Adquiridos e Licenciamento	17
Diagrama de Casos de Uso	18
Descrição de Casos de Uso	19
Buscar pesquisadores por palavra-chave [UC1]	19
Realizar busca lógica composta por palavra-chave com filtro [RF3]	20
Aplicar correção ortográfica das palavras chaves [RF9]	20
Calcular o indicador de relevância do pesquisador no tema Meio Ambiente [RF2]	21
Calcular indicador de relevância do pesquisador [UC2]	21
Registrar as palavras chaves [RF10]	21
Calcular indicador de relevância do pesquisador [UC2]	23
Realizar busca semântica [UC3]	23
Converter CV Lattes [UC4]	24
Converter dados de cadastro [UC5]	26
Converter dados disciplinas [UC6]	28
Gerar relatório palavras chaves [UC7]	29
Manter informações [UC8]	30
Manter palavras reservadas [UC9]	30
Matriz de Rastreabilidade	31
Detalhamento Técnico dos Requisitos Não-funcionais	32

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

Especificação dos Requisitos de Sistema

1. Introdução

Este documento tem por objetivo descrever os requisitos funcionais e elencar os requisitos não-funcionais e restrições da ferramenta Busc@NIMA. O desenvolvimento dessa ferramenta visa atender uma demanda do NIMA (Núcleo Integrado de Meio Ambiente) que está sendo atendida pelo grupo de pesquisa BioBD, sob a coordenação do professor Sérgio Lifschitz, do Departamento de Informática e também coordenador de pesquisas e membro do conselho do NIMA. Este projeto envolve as áreas de bancos de dados, recuperação de informações, web semântica e engenharia de software.

1.1.Finalidade

A ferramenta Busc@NIMA visa identificar pesquisadores da PUC-Rio cujos trabalhos de pesquisa, docência e/ou desenvolvimento envolvam a temática do meio ambiente. Através dessa ferramenta será possível divulgar as atividades de professores, alunos e funcionários nesta área para a sociedade em geral, considerando o especial interesse de outros pesquisadores assim como jornalistas.

A ideia básica consiste em indexar informações dos websites dos professores, informações do SAU online sobre disciplinas oferecidas, e também dos CV lattes da comunidade PUC-Rio e oferecer nomes e links de contato de professores, funcionários e alunos que estejam envolvidos com o tema, facilitando as buscas por parte da imprensa, da própria comunidade PUC-Rio de professores, alunos e funcionários administrativos e do público em geral. A ferramenta será disponibilizada via Web para acesso através do browser a partir do site institucional do NIMA.

1.2.Escopo

O desenvolvimento da ferramenta segue o ciclo de vida do Modelo Evolutivo uma vez que o escopo aberto permite que os requisitos da ferramenta sejam identificados gradativamente. Seguindo esse modelo, versões funcionais que atendam parcialmente a demanda geral, baseada nos requisitos conhecidos inicialmente, podem ser desenvolvidas e entregues. Nesse documento serão descritas as linhas gerais do que se espera das primeiras versões da ferramenta e novos requisitos poderão ser adicionados à medida que a ferramenta venha sendo divulgada e utilizada.

Em sua primeira versão o usuário poderá realizar uma busca sintática por palavra-chave (um único termo ou uma expressão) com filtros por tipo de recurso de interesse, como artigos, livros, capítulos, entre outros, na base de dados e o resultado das buscas inclui a lista de membros da comunidade PUC-Rio relacionados com o termo buscado, ordenado por um critério de relevância inicialmente fixo, definido de acordo com a quantidade de itens recuperados.

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

Na versões seguintes novas funcionalidades serão agregadas a ferramenta de busca, entre elas:

- a ordenação também levará em consideração um indicador que mede a similaridade entre as informações coletadas sobre o pesquisador e um conjunto de palavras-chave reservadas selecionadas por especialistas relativas à temática do meio ambiente.
- a busca contemplará operadores lógicos (OR, AND, NOT) para combinação de palavras-chave assim como a correção ortográfica automática desses termos.
- a ferramenta poderá fazer uso de buscas semânticas para identificar iniciativas na área do meio ambiente desvinculadas da sintaxe do conjunto de palavras reservadas.

A base de dados inicialmente será composta de informações extraídas do CV Lattes de professores, alunos e funcionários da PUC-Rio. Também serão contempladas informações sobre disciplinas extraídas do SAU online e dos websites dos professores.

A base de dados poderá ser posteriormente enriquecida com informações dos próprios pesquisadores, quando pelo título de um projeto ou nome de uma disciplina (ou outros recursos de interesse) não ficar claro que há relacionamento com o tema deste projeto.

Cabe ressaltar que a abordagem deste trabalho parte do pressuposto que a maior parte das informações sobre os membros da comunidade PUC-Rio está disponível publicamente. Entretanto, as próprias pessoas envolvidas ora não publicam suas atividades que envolvam o meio ambiente, ora publicam, mas sem explicitar que há relação com meio ambiente. Assim, caberá aos membros da PUC-Rio apenas o complemento e correção de informações não disponíveis publicamente.

1.3. Definições, Acrônimos e Abreviações

BioBD

Grupo de Pesquisa do Departamento de Informática focado em Banco de Dados e BioInformática

CNPQ

Conselho Nacional de Desenvolvimento Científico e Tecnológico

DTD

Document Type Definition (Documento para definição de tipos)

Padrão que descreve os elementos existentes em um documento XML, em termos de formato e estrutura, e permite que aplicativos compreendam a estrutura em uma árvore definida no documento XML.

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

ETL

Extract, Transform and Load (Extração, Transformação e Carga)

Processo de 3 etapas que visa copiar ou mover dados de uma fonte para outro repositório. Extração de dados consiste em se comunicar com fontes de dados para capturar os dados de interesse. A transformação pode realizar várias ações sobre os dados extraídos como padronização/harmonização/conversão (dados vindos de sistemas diferentes podem ter padrões diferentes seja de nomenclatura ou mesmo de tipos de dados), limpeza (remoção de dados que não são de interesse) e verificações de qualidade do conteúdo. A etapa final de carga visa o armazenamento dos dados resultantes no repositório de destino.

Lattes (Currículo e Plataforma)

A Plataforma Lattes, mantida pelo CNPQ, é responsável por armazenar e disponibilizar os dados do Currículo Lattes dos pesquisadores brasileiros. Atualmente é adotado pela maioria das instituições de fomento, universidades e institutos de pesquisa do País.

NIMA

Núcleo Integrado de Meio Ambiente

Ontologia

Os termos "vocabulário controlado" e "ontologia" podem ser usados de modo intercambiável na Ciência da Computação. A tendência é usar a palavra "ontologia" para uma coleção de termos mais complexa e possivelmente bastante formal, enquanto "vocabulário controlado" é usado quando esse formalismo estrito não é necessariamente usado ou apenas em um sentido muito amplo.

RDF

Resource Description Framework

Padrão para intercâmbio de dados na Web com recursos que facilitam a integração e interoperabilidade de dados, mesmo que os esquemas subjacentes sejam diferentes, e também suporta a evolução dos esquemas, sem exigir que todos os consumidores de dados sejam alterados.

Fonte: <https://www.w3.org/RDF/>

SAU

Sistema Acadêmico Universitário

Semantic Web

Web Semântica

Trata-se de um esforço colaborativo liderado pelo W3C com a participação de um grande número de pesquisadores e parceiros industriais que fornece uma estrutura comum, baseado no RDF, para permitir que os dados sejam compartilhados e reutilizados entre aplicativos, empresas e comunidades, com apoio de Vocabulários Controlados.

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

Fonte: <https://www.w3.org/2001/sw/>

Triplificação

É o processo de transformação de dados em qualquer formato estruturado ou semi estruturado (relational, XML, CSV, ...) para dados conectados em formato de triplas no modelo de grafo RDF.

Vocabulário Controlado

Define os conceitos e relacionamentos (também chamados de "termos") usados para descrever e representar um domínio ou área de interesse, para organizar o conhecimento, além de definir possíveis restrições ao uso desses termos. As técnicas de inferência na Web Semântica fazem uso desses vocabulários para resolver problemas na integração de dados (por exemplo, ambigüidades nos termos usados nos diferentes conjuntos de dados) ou para a descoberta de novos relacionamentos.

W3C

O World Wide Web Consortium (W3C) é uma comunidade internacional em que as organizações membros, uma equipe própria e o público em geral cooperam para desenvolver padrões da Web.

Fonte: <https://www.w3.org/Consortium/>

XML

eXtensible Markup Language.

XML é uma linguagem de marcação extensível de dados cujo principal objetivo é o intercâmbio de informações.

Fonte: <https://www.w3.org/XML/>

1.4. Organização do documento

No item 2.1 serão descritos os requisitos funcionais, inicialmente conhecidos, a serem atendidos pelo sistema. No item 2.2 são elencados os requisitos não funcionais a serem aplicados ao sistema e ao ambiente onde o sistema opera. No item 2.3 são descritas as características e restrições associadas ao processo de desenvolvimento do sistema e no item 2.4 são registradas questões sobre licenciamento.

No item 3 o Diagrama de Casos de Uso do sistema é apresentado e no item 4 os casos de uso são detalhados em seus fluxos principais e alternativos. O item 5 apresenta a Matriz de Rastreabilidade entre requisitos funcionais e os casos de uso.

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

2. Requisitos

2.1. Funcionalidades

2.1.1. O sistema deve permitir que o usuário realize uma busca sintática simples por palavra-chave com filtro [RF1]

A ferramenta de busca deve exibir uma página inicial onde o usuário pode inserir uma palavra chave ou uma expressão, vinculadas ou não ao tema de meio ambiente, e selecionar um conjunto de elementos de interesse para a busca. Com a entrada do usuário o sistema deve buscar em campos do tipo texto existentes na base de dados, vinculados aos elementos de interesse selecionados, se estes contêm a palavra chave ou expressão utilizada pelo usuário em seu conteúdo. Os resultados encontrados, devem ser contados e agrupados por tipo de elemento, e associados aos dados de contato e biografia do pesquisador, que pode ser um professor ou um aluno da PUC-Rio.

2.1.2. O sistema deve realizar o cálculo de indicador de relevância do pesquisador no tema Meio Ambiente [RF2]

Cada pesquisador deve ter seu Currículo Lattes avaliado em relação a sua vinculação com o tema Meio Ambiente. O indicador de relevância deve ser calculado a partir da similaridade entre os termos usados para descrever a produção e atividade acadêmica do pesquisador e um conjunto de palavras chaves reservadas associadas ao tema Meio Ambiente. O conjunto de palavras chaves inicial é apresentado abaixo.

Sustentabilidade	Meio-ambiente	Ambiental
Paisagem	Natureza	Poluição
Ecologia	Resiliência	Vulnerabilidade
Água	Biodiversidade	Resíduos
Energia	Mudança climática	Mobilidade
Educação ambiental	Saúde	Espaço de convivência
Impacto ambiental	Reciclagem (materiais)	

Esse indicador deverá ser recalculado sempre que uma nova versão do Currículo Lattes for carregada na base de dados ou se o conjunto de palavras chaves for alterado.

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

2.1.3.O sistema deve permitir que o usuário realize uma busca lógica composta por palavra-chave com filtro [RF3]

TBD

2.1.4.O sistema deve permitir que o usuário realize uma busca semântica apoiada por Ontologia de Meio Ambiente [RF4]

TBD

2.1.5.O sistema deve realizar a conversão e carga de Currículo Lattes em base de dados triplicada [RF5]

O CNPq disponibiliza, através da Plataforma Lattes, às instituições a possibilidade de extração de dados da base de Currículos Lattes por meio de uma ferramenta chamada Lattes Extrator. Essa ferramenta gera um arquivo no formato XML para cada Currículo Lattes contendo os dados cadastrados pelo pesquisador. Esse arquivo XML segue o padrão de nomenclatura <ID>.xml onde o ID é o identificador único de x dígitos do pesquisador na plataforma. Por exemplo, para o professor Sérgio Lifschitz o ID é 8164403687403639 e o arquivo correspondente é o 8164403687403639.xml. A estrutura do arquivo segue o esquema DTD descrito em <http://impl.cnpq.br/impl/Gramaticas/Curriculo/DTD/Documentacao/DTDCurriculo.pdf>

O conjunto de arquivos xml correspondentes aos pesquisadores (alunos e professores) vinculados à PUC-Rio, será gerado através do Lattes Extrator e disponibilizado periodicamente (a cada 2 meses) .

A conversão do formato XML para o modelo de grafo RDF deve ser realizada com o suporte de uma ferramenta de ETL e com scripts/utilitários padronizados que preservem o conteúdo e a relação entre os elementos.

2.1.6.O sistema deve realizar a conversão e carga de arquivo CSV com dados de cadastro de usuários em base de dados triplicada [RF6]

A área tal (qual área da PUC faz isso?) deve disponibilizar periodicamente (qual periodicidade?) um arquivo em formato CSV contendo dados de cadastro de professores e funcionários vinculados à PUC-Rio extraídos dos sistemas acadêmicos (qual ou quais?). A estrutura do arquivo deverá conter no mínimo as seguintes colunas: matricula, email, usuário (nome) e home_page.

A conversão do formato CSV para o modelo de grafo RDF deve ser realizado com o suporte de uma ferramenta de ETL e com scripts/utilitários padronizados que preservem o conteúdo e a relação entre os elementos.

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

2.1.7.O sistema deve realizar a conversão e carga de planilha Excel com dados de disciplinas do SAU em base de dados triplicada [RF7]

A equipe do NIMA deve disponibilizar semestralmente uma planilha Excel contendo dados de disciplinas, cursos e professores extraídos do SAU, separados por grupos (graduação, pós graduação, extensão, ...) e vinculados ao conjunto de palavras chaves inicial elencado no RF2. A estrutura do arquivo deverá conter no mínimo as seguintes colunas: nome da disciplina, descrição/ementa, nome do professor.

A conversão de cada aba da planilha Excel para o formato CSV, assim como do formato CSV para o modelo de grafo RDF deverão ser realizadas com o suporte de uma ferramenta de ETL e com scripts/utilitários padronizados que preservem o conteúdo e a relação entre os elementos.

2.1.8.O sistema deve realizar a triplicação das fontes de dados utilizando ontologias de domínio para atender aos padrões da Web Semântica [RF8]

A tripla RDF gerada pela transformação corresponde ao padrão <S, P, O>, onde:

- S é um URIref, chamado de sujeito da declaração,
- P é um URIref, chamado de propriedade ou predicado que denota um relacionamento binário e
- O é um URIref ou um literal, chamado o objeto da instrução; se O é um literal, também é chamado de valor da propriedade P

A notação de grafos RDF converte um conjunto de triplas RDF em um grafo direcionado, com nós representando sujeitos ou objetos e arcos representando propriedades. Os nós são rotulados por meio de URIs que descrevem recursos ou literais (ou seja, cadeias de caracteres ou números) ou são não rotulados, chamados nós em branco.

Durante a etapa de transformação das fontes de dados devem ser utilizadas ontologias de domínio padronizadas, conhecidas e publicamente disponíveis que tenham sido construídas utilizando os padrões tecnológicos da Web Semântica. A utilização visa estabelecer uma relação organizada e padronizada entre termos, favorecendo a possibilidade de contextualização dos dados, extraídos de diversas fontes, e facilitando o processo de interpretação dos dados pela ferramenta de busca. Os recursos (sujeitos e objetos) devem corresponder a conceitos presentes na ontologias utilizadas assim como os predicados aos relacionamentos e atributos. A seguir está a lista inicial de ontologias utilizadas:

Ontologia	URIref	short name
Bibliographic Ontology Specification	http://purl.org/ontology/bibo/	bibo

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

BIO: A vocabulary for biographical information	http://purl.org/vocab/bio/0.1/	bio
Hierarquia Área de Conhecimento - Especialidades do CNPQ	http://estatico.cnpq.br/bi/CNPQ/DadosAbertos/Tabelas/AreaConhecimento/area_conhecimento.xsd#	cnpqterms
Dublin Core Metadata Initiative (DCMI)	http://purl.org/dc/elements/1.1/	dc
DCMI Metadata Terms	http://purl.org/dc/terms/	dcterms
Friend of a Friend (networks)	http://xmlns.com/foaf/0.1/	foaf
An ontology for representing scholarship	https://duraspace.org/vivo/	vivo
Curriculum Course Syllabus Ontology (vide figura 1)	https://w3id.org/ccso/ccso#	ccso

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

2.1.13.O sistema deve permitir ao administrador manter a lista de palavras chaves reservadas associadas ao tema Meio Ambiente [RF13]

TBD

2.2. Características e restrições do produto

<u>Segurança</u>
[RNF1] Restringir a troca de informações via browser do usuário
[RNF2] Apenas a equipe de desenvolvimento pode acessar os servidores e software backend do produto
[RNF3] Restringir as permissões de acesso aos componentes/ambientes com perfil adequado aos usuários e funções
[RNF4] Proteger a entrada de dados da ferramenta de consulta de ataques via Internet
<u>Usabilidade</u>
[RNF5] Uso de design responsivo na interface gráfica da ferramenta de busca
[RNF6] Utilizar hints nos campos e botões de navegação da ferramenta de busca
[RNF7] Utilizar logotipo da PUC e do NIMA na parte superior da ferramenta de busca
<u>Confiabilidade</u>
[RNF8] Configurar os componentes da ferramenta de busca para restart automático em caso de queda da VM ou outros problema de infra estrutura
[RNF9] Criar página de erro padrão para ferramenta de busca
[RNF10] Realizar backup periódico das máquinas virtuais que atendem ao projeto, contemplando principalmente a base de dados e os arquivos de carga
<u>Desempenho</u>
[RNF11] Exibir o resultado da consulta na ferramenta de busca em no máximo 2 minutos
[RNF12] Suportar até 5 usuários simultâneos na ferramenta de busca sem degradação no tempo de resposta
<u>Portabilidade / Compatibilidade</u>

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

[RNF13] Suportar o uso da ferramenta de busca nos principais browsers de dispositivos com sistema operacional Windows, Linux e Android
[RNF14] Suportar a carga de dados a partir de arquivos em diferentes formatos como XML, RDF, banco de dados, CSV e planilhas Excel
<u>Disponibilidade</u>
[RNF15] O sistema deve estar disponível 99% do tempo.
<u>Dados e Armazenamento</u>
[RNF16] Suportar modelo de dados flexível
[RNF17] Suportar indexação de dados em formato texto

2.3.Características e restrições do processo

No que diz respeito a ferramentas e ambientes usados para o desenvolvimento do sistema, principalmente da ferramenta de busca, o projeto seguirá o padrão adotado pelos demais projetos do laboratório BioBD: linguagem python 3 da distribuição Anaconda, framework Flask, controle de versão com o Git, sistema operacional Ubuntu nos servidores virtuais e Windows nas máquinas desktop dos desenvolvedores.

A utilização da linguagem **python** para o desenvolvimento da ferramenta de busca está alinhada aos requisitos da aplicação uma vez que é uma linguagem de programação multiparadigma, ou seja, suporta diversos paradigmas de desenvolvimento, como, orientado à objetos, funcional, imperativo, interpretado, entre outros, multiplataforma e open source. Além disso trata se de uma linguagem, adotada nas disciplinas de graduação da PUC, devido a sua curva de aprendizagem facilitada e ter se tornado uma das principais linguagem para o desenvolvimento Web, IA, Machine Learning e Big Data, áreas em grande crescimento nos últimos anos.

Anaconda é uma distribuição de Python e R que simplifica o gerenciamento de pacotes e ambientes virtuais de um projeto. O Anaconda será utilizado para o gerenciamento do projeto por meio de um ambiente virtual Python isolado, que separa as dependências específicas do projeto em uma única “pasta”, mantendo todos os pacotes e configurações específicas do projeto em um único lugar no computador.

Para o desenvolvimento web, os desenvolvedores contam com dois importantes microframeworks, o Django e **Flask**. O grupo de desenvolvedores do BioBD adotou o Flask

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

devido a necessidade de rápida prototipação de aplicações web do tipo CRUD e desenvolveu um tutorial (documento “***Tutorial Flask-BioBD***”) para orientar quanto aos primeiros passos com o uso da tecnologia. Flask é baseado nos pacotes Werkzeug e Jinja2 que oferece suporte no gerenciamento de requests ao site (Werkzeug) e na geração das páginas HTML (Jinja2). O Flask oferece flexibilidade quanto à características como persistência de dados e sistema de login também trabalha com componentes “plugáveis” (extensões).

Quanto ao banco de dados, o projeto fará uso do **AllegroGraph**, um banco NoSQL e o acesso a esse banco pela ferramenta de busca fará uso de uma API em python. Trata-se de um banco multi-modelo (Documento em JSON, JSON-LD e Grafo em RDF) que atende aos padrões W3C para a Web Semântica. Entre outras linguagens, AllegroGraph oferece suporte a SPARQL que é uma linguagem de consulta padrão para dados linkados. O software também suporta de modo nativo a indexação de texto livre nos objetos de triplas cujos predicados foram registrados para indexação. Essa solução atende aos requisitos RNF16 e RNF17.

A escolha do AllegroGraph se deu por ser um Triple Store que permite o maior número de triplas por collection em uma versão sem custo em relação às demais opções disponíveis. A experiência dos professores Jefferson de Barros Santos e Edward Hermann Haeusler em outro projeto que envolvia a triplificação dos dados do CV Lattes com esse Triple Store e com o processo de conversão de XML em RDF permitiu uma rápida prototipação da ferramenta de busca.

Uma ferramenta de ETL será usada para o atendimento dos requisitos RF5, RF6, RF7 e RF8. Esse tipo de ferramenta em sua maioria já possui funções específicas para a criação de processos automatizados de conversão e carga de dados, não requer conhecimento de linguagens de programação de baixo nível, fazem uso de 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 e permitem a conexão a repositórios remotos (como fonte ou destino dos dados). A ferramenta de suporte ao processo de ETL escolhidas foi a **Linked Pipes**, que foi desenvolvida para manipular prioritariamente dados em formato de tripla.

Anaconda	https://www.anaconda.com/products/individual
Flask	https://docs.google.com/document/d/1OY2AodrTA7CXE9ZM2OLi3ADtsEYxzTGoxqHLikaca2s/edit?ts=5ec80e4d#
Git	https://github.com/sergiolif/biobd-nima
Allegro Graph	https://franz.com/agraph/support/documentation/6.4.2/agraph-introduction.html https://franz.com/agraph/support/documentation/6.4.2/python/tutorial.html
Linked Pipes	https://etl.linkedpipes.com/

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

O acompanhamento do desenvolvimento das atividades e comunicação entre os membros da equipe ocorria através de reuniões periódicas, inicialmente presenciais e semanais mas em função da pandemia do coronavírus passaram a ser virtuais. Também adotou-se como prática nos projetos do BioBD o uso de uma ferramenta de Blog para registro de atividades, decisões, conhecimentos e procedimentos no escopo de cada projeto, visando a gestão do conhecimento.

2.4.Componentes Adquiridos e Licenciamento

O projeto NIMA não prevê a aquisição de softwares específicos.

A infraestrutura de hardware e software básico para os servidores está sendo fornecida pelos serviços da Cloud do DI.

As ferramentas de desenvolvimento utilizadas pela equipe do BioBD são distribuídas como software livre, sem custo de uso.

A versão Free do Allegro Graph que está em uso não tem data de expiração mas impõem a restrição de no máximo 5 milhões de tuplas em cada repositório. Essa restrição irá implicar em ajustes no projeto físico como a divisão em múltiplos repositórios no mesmo servidor.

De modo a subsidiar decisões futuras sobre questões que envolvem custos de aquisição e pagamentos por serviços foi realizado um levantamento sobre as formas de licenciamento do AllegroGraph:

Existem 3 distribuições para instalação em servidores próprios: Free (limite de 5 Milhões de triplas por repositório, máximo de 3 servidores em modo distribuído), Developer (limite de 50 Milhões de triplas por repositório, máximo de 4 cores em um único servidor) e Enterprise (sem limites). A primeira é gratuita e as demais o valor deve ser obtido através de cotação por e-mail. É importante ressaltar que de acordo com o Termo de Licença de Uso a aquisição não garante atualização de versão posterior.

Também é possível contratar como por serviço na AWS (Amazon). Na opção Free só é cobrada a infraestrutura da máquina (\$0.159 por hora + \$0.19 por GB-mês de armazenamento). Na versão Enterprise é cobrado tanto a licença quanto a infraestrutura (\$1.406 por hora + \$0.19 por GB-mês de armazenamento) na configuração inicial de 16G de RAM e 4 Cores recomendada pelo fornecedor.

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

3.Diagrama de Casos de Uso

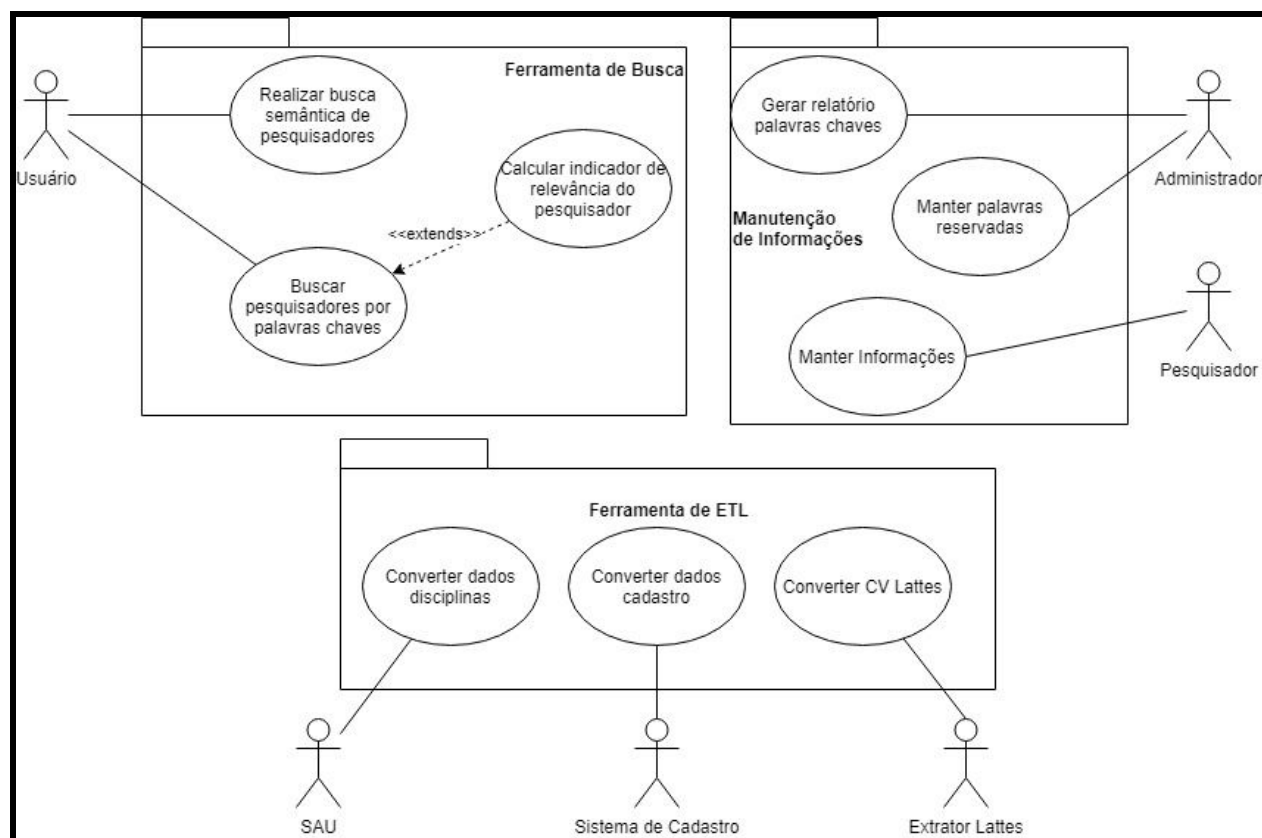


Figura 2 - Diagrama de Casos de Uso do Sistema NIMA

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

4. Descrição de Casos de Uso

4.1. Buscar pesquisadores por palavra-chave [UC1]

Nome	Buscar pesquisadores por palavra-chave	
Objetivo	Permitir ao ator realizar busca de pesquisadores por palavra-chave	
Requisitos Funcionais	[RF1], [RF3], [RF9], [RF10]	
Atores	Usuário (Público)	
Prioridade	Alta	
Frequência de uso	Eventual	
Criticalidade	Alta	
Pré condições	Não se aplica	
Condição de Entrada	Usuário deve selecionar a opção X no site do NIMA na Web.	
Fluxo Principal	Ator	Sistema
	1. Informa palavra(s) chave e seleciona os tipos de elementos de interesse na busca [RN1][A1][E1][E2][T1]. 2. Clica em Buscar.	
		3. Registrar as palavras chaves [RF10] 4. Busca por triplas onde o objeto do tipo de dados texto contenha a(s) palavra(s) chave [RN2][E3] 5. Exibe o nome do autor, seu tipo e a quantidade de ocorrências ordenado pelos dois últimos campo e com paginação padrão de 10 itens [T2].
	6. Seleciona um pesquisador para detalhamento [A2]	

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

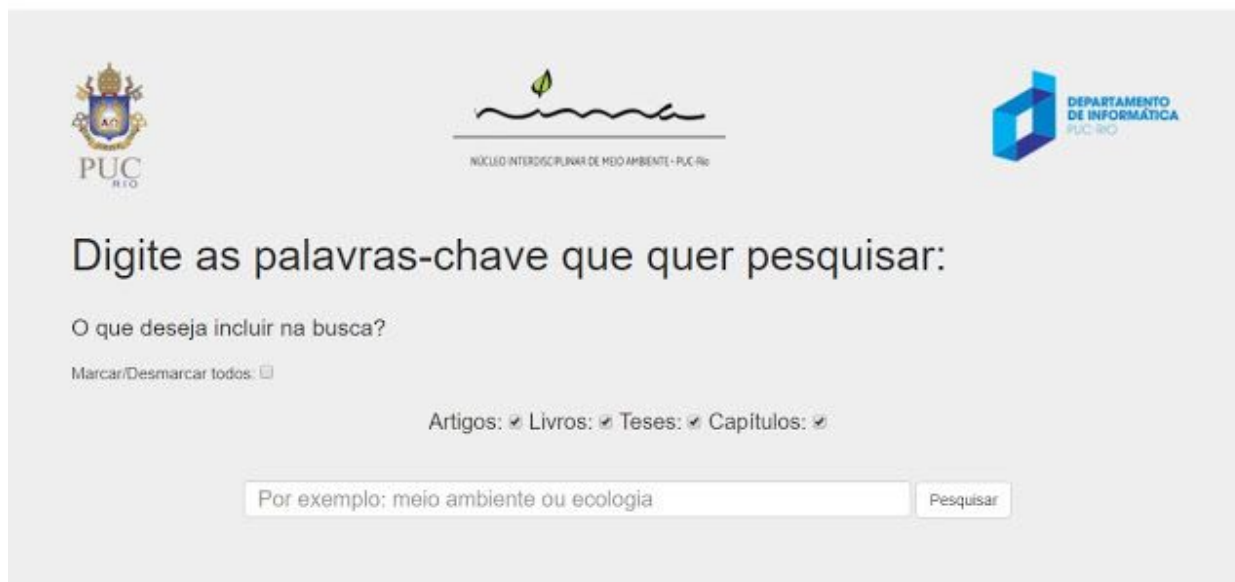


		7. Exibe detalhes sobre o pesquisador selecionado como biografia, e-mail, site, link para o CV Lattes completo e sobre os itens de interesse marcados que correspondem ao critério de busca.[T4]
	8. Clica no link do CV Lattes (FIM)	
Fluxo Alternativo	Ator	Sistema
	[A1] Informa palavra(s) chaves com operadores lógicos para montar a string de busca (OR, AND, NOT)	
		<i>Realizar busca lógica composta por palavra-chave com filtro [RF3]</i>
	[A2] Filtra por nome do pesquisador, altera a quantidade de itens por página ou altera o critério de ordenação. [T3]	
		Modifica os dados apresentados na tela conforme a nova configuração
Fluxo de Exceção	Ator	Sistema
	[E1] O usuário não digitou uma palavra	
		Exibir mensagem "Preencha esse campo"
	[E2] O usuário digitou palavra(s) com erro ortográfico	
		<i>Aplicar correção ortográfica das palavras chaves [RF9]</i>
	[E3] Não existem elementos que atendam ao critério de busca	

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

	estabelecido	
		Exibir tela padrão de elementos não encontrados [T5]
Pontos de Extensão	<i>Calcular o indicador de relevância do pesquisador no tema Meio Ambiente [RF2]</i> <i>Calcular indicador de relevância do pesquisador [UC2]</i>	
Pós condições	<i>Registrar as palavras chaves [RF10]</i>	
Regras de negócio	[RN1] Elementos de interesse são: artigos, teses, livros e capítulos. Novos elementos poderão ser incluídos à medida que os dados forem carregados. Por padrão todos os elementos de interesse não devem estar marcados ao carregar a tela. [RN2] A pesquisa deve ser feita no título dos elementos de interesse, na biografia e no nome dos pesquisadores. Novos atributos do tipo texto poderão ser incluídos à medida que os dados forem carregados.	

Telas

T1

Digite as palavras-chave que quer pesquisar:

O que deseja incluir na busca?

Marcar/Desmarcar todos: ☐

Artigos: ☒ Livros: ☒ Teses: ☒ Capítulos: ☒

Por exemplo: meio ambiente ou ecologia

T2 & T3

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

Artigos: ☒ Livros: ☒ Teses: ☒ Capítulos: ☒

Piracicaba

Temos **7** autores relacionados com a(s) palavra(s) **Piracicaba** :

Mostrar pessoas por página Filtrar:

Nome	Tipo	Ocorrências
R. VICTORIA	Professor	1
M. H. B. FALÓTICO	Professor	1
L.A. MARTINELLI	Professor	1
JOSE MARCUS DE OLIVEIRA GODOY	Professor	1
EPAMINONDAS S. B. FERRAZ	Professor	1
ANNE HELENE FORSTIER	Professor	1
PAULO CESAR TEIXEIRA	Professor	1

T4

ALLAN NOGUEIRA DE ALBUQUERQUE
Termo pesquisado: 'robóticos'

BIOGRAFIA

* É Doutor em Engenharia Mecânica pela PUC-Rio (2017) e possui Mestrado (2012) e Graduação (2009) em Engenharia Mecânica na mesma instituição. Já atuou na área de inspeção veicular e realizou atividades relacionadas à calibração de transdutores de força e ensaios mecânicos. Participou de projetos do Centro de Avaliação Não Destrutiva (CAND) no desenvolvimento de veículos robóticos de inspeção. Atualmente, trabalha como Engenheiro de Pesquisa no ITUC/PUC-Rio, atuando na concepção, testes e validação de sistemas robóticos de inspeção visual, mecanismos articulados para reprodução de movimentos e equipamentos para o Setor de Petróleo e Gás, além de ser Professor Agregado na PUC-Rio.

SITE

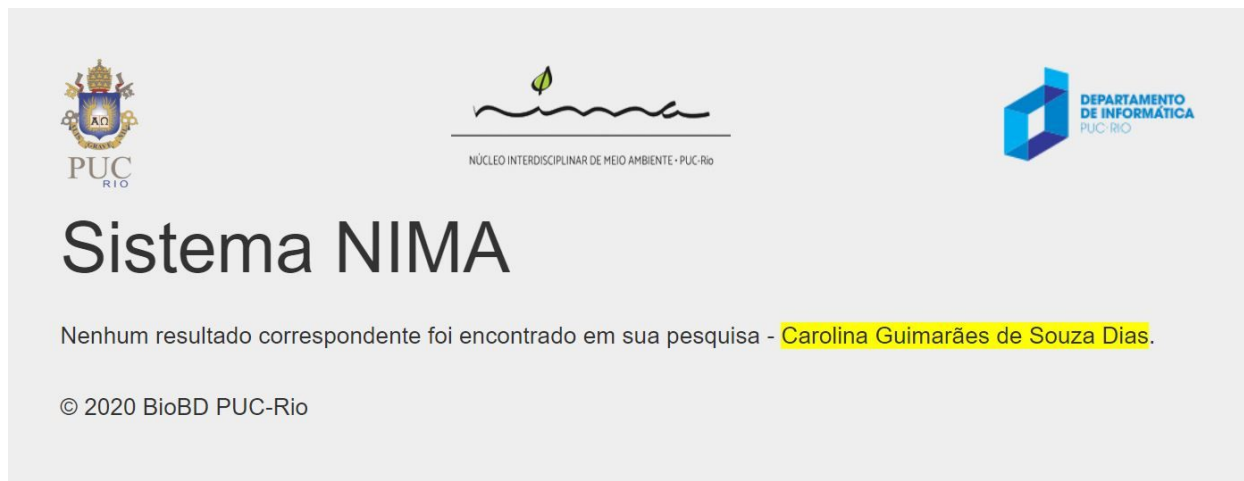
* www.ituc.puc-rio.br
* www.ituc.puc-rio.br

EMAIL

* allan@puc-rio.br
* allanalbuquerque@gmail.com

T5

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	



4.2. Calcular indicador de relevância do pesquisador [UC2]

TBD

4.3. Realizar busca semântica [UC3]

TBD

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

4.4.Converter CV Lattes [UC4]

Nome	Converter CV Lattes	
Objetivo	Carregar dados gerados pelo ator na base de dados do sistema	
Requisitos Funcionais	[RF5], [RF8]	
Atores	Extrator Lattes	
Prioridade	Alta	
Frequência de uso	Bimestral	
Criticalidade	Alta	
Pré condições	Executar a aplicação de Extração dos CVs Lattes na plataforma Lattes do CNPQ	
Condição de Entrada	Arquivos XML com os CVs Lattes disponibilizados no servidor	
Fluxo Principal	Ator	Sistema
	1. Extrair CVs Lattes [RN3] 2. Copiar conjunto de arquivos XML para o servidor de ETL.	
		3. Ler arquivos XML no diretório de entrada 4. Verificar se os arquivos XML estão bem formados [E1] 5. Converter os arquivos XML em RDF [E2] 6. Remover triplas desnecessárias e duplicadas (caso existam) 7. Mapear conceitos e relacionamentos em Ontologias 8. Gravar arquivos RDF no diretório de saída 9. Copiar conjunto de arquivos RDF

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

		para o servidor de banco de dados. 10. Carregar os arquivos nos repositórios [E3] (FIM)
Fluxo Alternativo	Ator	Sistema
Fluxo de Exceção	Ator	Sistema
		[E1] Existem problemas de formatação nos arquivos XML que impedem o seu processamento
	Voltar ao passo 1	
		[E2] Existem problemas de conteúdo nos arquivos XML que impedem o seu processamento
	Voltar ao passo 1	
		[E3] Existem problemas nos arquivos RDF que impedem a carga
		Voltar ao passo 3
Pontos de Inclusão	Não há	
Pós condições	Dados carregados na base	
Regras de negócio	[RN1] Os arquivos XML devem conter somente CV Lattes de pesquisadores (alunos, professores, funcionários) vinculados à PUC-Rio.	

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

4.5. Converter dados de cadastro [UC5]

Nome	Converter dados de cadastro	
Objetivo	Carregar dados gerados pelo ator na base de dados do sistema	
Requisitos Funcionais	[RF6], [RF8]	
Atores	Sistema de cadastro	
Prioridade	Média	
Frequência de uso	Sob demanda	
Criticalidade	Média	
Pré condições	Gerar arquivo CSV com informações extraídas do Sistema de Cadastro de Funcionários da PUC-RIO	
Condição de Entrada	Arquivos CSV com dados de cadastro disponibilizados no servidor	
Fluxo Principal	Ator	Sistema
	1. Gerar arquivo CSV [RN1] 2. Copiar arquivo para o servidor de ETL.	
		3. Ler arquivos CSV no diretório de entrada 4. Converter CSV em RDF [E1] 5. Remover triplas desnecessárias e duplicadas (caso existam) 6. Mapear matrícula em ID Lattes pelo nome [E2] 7. Mapear conceitos e relacionamentos em Ontologias 8. Gravar arquivo RDF no diretório de saída 9. Copiar arquivo RDF para o servidor de banco de dados.

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

		10. Carregar arquivo no repositório [E3] (FIM)
Fluxo Alternativo	Ator	Sistema
Fluxo de Exceção	Ator	Sistema
		[E1] Existem problemas de formatação ou conteúdo no arquivo CSV que impede o seu processamento
	Voltar ao passo 1	
		[E2] Não foi encontrado CV Lattes para o funcionário com base no nome de cadastro
		Gerar log de nomes não encontrados Continua no passo 7
		[E3] Existem problemas nos arquivos RDF que impedem a carga
		Voltar ao passo 3
Pontos de Inclusão	Não há	
Pós condições	Dados carregados na base	
Regras de negócio	[RN1] O arquivo CSV deve conter uma linha para cada pesquisador (professores, funcionários) vinculados à PUC-Rio com pelo menos matrícula, nome, e-mail e site.	

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

4.6. Converter dados disciplinas [UC6]

Nome	Converter dados de disciplinas	
Objetivo	Carregar dados gerados pelo ator na base de dados do sistema	
Requisitos Funcionais	[RF6], [RF8]	
Atores	SAU	
Prioridade	Média	
Frequência de uso	Semestral	
Criticalidade	Média	
Pré condições	Gerar planilha Excel com informações extraídas de disciplinas e cursos vinculados ao Meio Ambiente do SAU da PUC-RIO	
Condição de Entrada	Planilha Excel com dados de disciplinas e cursos disponibilizadas no servidor	
Fluxo Principal	Ator	Sistema
	1. Gerar arquivo Excel [RN1] 2. Copiar arquivo para o servidor de ETL.	
		3. Ler arquivo Excel no diretório de entrada 4. Converter arquivo Excel para CSV (um CSV por aba) 5. Converter CSV em RDF [E1] 6. Remover triplas desnecessárias e duplicadas (caso existam) 7. Mapear professor da disciplina em ID Lattes pelo nome do professor [E2] 8. Mapear conceitos e relacionamentos em Ontologias 9. Gravar arquivos RDF no diretório de saída

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

		10. Copiar arquivo RDF para o servidor de banco de dados. 11. Carregar arquivo no repositório [E3] (FIM)
Fluxo Alternativo	Ator	Sistema
Fluxo de Exceção	Ator	Sistema
		[E1] Existem problemas de formatação ou conteúdo no arquivo CSV que impede o seu processamento
	Voltar ao passo 1	
		[E2] Não foi encontrado CV Lattes para o professor com base no nome de cadastro
		Gerar log de nomes não encontrados Continua no passo 8
		[E3] Existem problemas nos arquivos RDF que impedem a carga
		Voltar ao passo 3
Pontos de Inclusão	Não há	
Pós condições	Dados carregados na base	
Regras de negócio	[RN1] A planilha Excel deve conter uma aba para cada tipo de curso (graduação, pós graduação, extensão, ...) e uma linha para cada disciplina com pelo menos nome, curso, ementa e professor.	

4.7. Gerar relatório palavras chaves [UC7]

TBD

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

4.8. Manter informações [UC8]

TBD

4.9. Manter palavras reservadas [UC9]

TBD

Busc@NIMA	Version: 1.3
Especificação dos Requisitos de Sistema	Date: 07/07/2020
ERS_BuscaNIMA	

5. Matriz de Rastreabilidade

	Caso de Uso								
RF	1	2	3	4	5	6	7	8	9
1	v1								
2		v2							
3	v3								
4			v4						
5				v1					
6					v1				
7						v1			
8				v1	v1	v1			
9	v3								
10	v2								
11							v2		
12								v5	
13									v2