UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

LUAM LEIVERTON PEREIRA DOS SANTOS

INTEGRAÇÃO DE DADOS PARA APOIO À TOMADA DE DECISÃO NO CONTEXTO DA AUTO-AVALIAÇÃO DE PROGRAMAS DE PÓS-GRADUAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO DE ESPECIALIZAÇÃO

DOIS VIZINHOS
2022

LUAM LEIVERTON PEREIRA DOS SANTOS

INTEGRAÇÃO DE DADOS PARA APOIO À TOMADA DE DECISÃO NO CONTEXTO DA AUTO-AVALIAÇÃO DE PROGRAMAS DE PÓS-GRADUAÇÃO

DATA INTEGRATION TO SUPPORT DECISION MAKING IN SELF-EVALUATION OF POSTGRADUATE PROGRAMS

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Marcelo Teixeira

DOIS VIZINHOS
2022

LUAM LEIVERTON PEREIRA DOS SANTOS

INTEGRAÇÃO DE DADOS PARA APOIO À TOMADA DE DECISÃO NO CONTEXTO DA AUTO-AVALIAÇÃO DE PROGRAMAS DE PÓS-GRADUAÇÃO

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Data de aprovação: 20/setembro/2022

Marcelo Teixeira
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

Dalcimar Casanova
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

Jefferson Tales Oliva
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

DOIS VIZINHOS
2022

AGRADECIMENTOS

Meus sinceros agradecimentos são remetidos à Deus, pai todo poderoso que me guia ao aperfeiçoamento pessoal e profissional contínuos.

Aos meus pilares de sustentação, minha amada esposa Leandra Rodrigues Vilela Pereira e a minha amada filha Luany Helena Pereira Vilela, pelo apoio irrestrito, dedicação, força e confiança para seguir visando aperfeiçoar academicamente e profissionalmente em busca do progresso da nossa família.

Aos meus pais, Maria dos Santos Pereira e Luiz Pereira dos Santos, pela origem, ensinamentos e educação investidos, assim como aos meus irmãos e sobrinhos.

Aos colegas de curso, pelo conhecimento compartilhado e atuação colaborativa na consecução dos projetos integradores ao longo da formação.

A UTFPR pelo excelente projeto de curso, recursos e infraestrutura. Aos docentes pela generosidade do conhecimento compartilhado, pelo tempo dedicado, pela base educacional disponibilizada. Em especial, ao meu orientador Prof. Dr. Marcelo Teixeira, por todo apoio a mim dedicado, pelas intervenções e sugestões, e por ter permanecido tão presente e atuante ao longo do desenvolvimento deste trabalho.

RESUMO

Uma das etapas mais importantes e críticas do processo educacional em Cursos de Pós Graduação em Instituições Federais de Ensino Superior se refere à Avaliação dos Programas. Nesta etapa, os resultados obtidos são confrontados com indicadores baseados nos objetivos e metas de desenvolvimento estratégico educacional, tornando possível a verificação de qualidade dos cursos e a necessidade de implementação de ações para correção de rumos ou consolidação das práticas já aplicadas visando a melhoria contínua dos cursos deste segmento. Neste contexto, este trabalho tem como objetivo estabelecer um modelo de integração de dados com especificação e implementação de um sistema de A autoavaliação para cursos de Pós Graduação. O Sistema resultante é modularizado em cinco componentes que agregam funcionalidades que ambicionam auxiliar à tomada de decisão reunindo dados de múltiplas fontes e minimizando esforço e custo, além de viabilizar o registro de informação temporal com definição e obtenção automática de indicadores relevantes para controle e monitoramento da qualidade dos cursos.

Palavras-chave: Integração de dados; Séries Temporais; Avaliação de cursos; Pesquisa e Pós Graduação.

ABSTRACT

One of the most importants and critical stages of the educational process in PostGraduate Courses in Federal Education Institutions refers to the Program's Evaluation. In this stage, the obtained results are confronted with indicators based on objectives and goals of strategic educational development, making possible to verify the courses quality and the action's implementation necessity to correct directions or consolidate already applied pratices, seeking the countinuous courses improvement in this segment. In this context, this work aims to establish a data integration model with specification and implementation of a Self Assessment System for Postgraduate Courses. The resulting system is modularized into five components that add functionality that aim to support decision making by gathering data from multiple sources and minimizing effort and cost, in addition to enabling the recording of temporal informations with definition and obtaining automatic creation of relevant indicators to control and monitor the courses quality.

Keywords: Data Integration; Time Series; Courses Evaluation; Research and PostGraduation.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
API	Application Programming Interface
BPM	Business Process Management
BI	Business Intelligence
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPQ	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CART	Classification and Regression Tree
CV	Currículum Vitae
DIA	Dia Diagram Editor
ELT	Extraction, Load, Transform
EM	Expectation–Maximization Algorithm
FIOCRUZ	Fundação Osvaldo Cruz
HTML	Hipertext Markup Language
IA	Inteligência Artificial
IBM	International Business Machines Corporation
ORM	Object-Relational Mapping
OWL	Ontology Web Language
JSON	JavaScript Object Notation
MEC	Ministério da Educação
OLAP	Online Analytical Processing
REST	Representational State Transfer
SAS	Software de Analytics & Soluções
SIGA	Sistema de Gestão Acadêmica
SIGRH	Sistema de Gestão de Recursos Humanos

SQL	Structured Query Language
SOAP	Simple Object Access Protocol
SNPG	Sistema Nacional de Pós Graduação
Sucupira	Sistema Unificado de Currículos e Programas: Identificação de Redes Acadêmicas
UFPE	Universidade Federal de Pernambuco
Univasf	Universidade Federal do Vale do São Francisco
Univile	Universidade da Região de Joinville
WSDL	Web Service Description Language
XML	eXtensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

Uma das etapas mais importantes e críticas do processo educacional em Cursos de Pós Graduação em Instituições Federais de Ensino Superior se refere à Avaliação dos Programas. Nesta etapa, os resultados obtidos são confrontados com indicadores baseados nos objetivos e metas de desenvolvimento estratégico educacional, tornando possível a verificação de qualidade dos cursos e a necessidade de implementação de ações para correção de rumos ou consolidação das práticas já aplicadas visando a melhoria contínua dos cursos deste segmento.

Na Pós Graduação, um dos parâmetros mais importantes que servem de referência para esta avaliação diz respeito à produção científica qualificada dos pesquisadores (docentes e discentes) vinculados aos programas, mensurado pela capacidade de validação dos resultados obtidos por seus pares, através da submissão dos resultados e posterior publicação, difusão e impacto das pesquisas, visando o progresso da Ciência e agregando valor à sociedade.

No modelo atual de Avaliação, as coordenações de curso enviam dados continuamente para apreciação e custódia do Ministério da Educação, sob responsabilidade da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-CAPES, responsável pelo reconhecimento e manutenção dos cursos, para tal é disponibilizada na Plataforma Sucupira um sistema de Coleta de Dados, onde as informações são inseridas ou importadas (CAPES, 2020b). Para uma eficiente alimentação do 'Coleta', faz-se necessário um processo de recuperação da informação eficiente, reunindo dados de várias fontes e tornando possível um acompanhamento ostensivo e institucional do progresso ou regressão nos cursos e seu nível de qualidade pelas Instituições de Ensino, de modo a prover uma intervenção efetiva para correção de rumos.

1.1 Problema de Pesquisa

Acompanhar a produção qualificada dos pesquisadores, de modo a otimizar resultados pode ser considerada uma tarefa complexa e custosa, porque algumas análises dependem da disponibilidade de um grande volume de dados, que comumente são registrados e armazenados em diferentes meios de armazenamento sem acesso integrado, ou ainda sob formatos heterogêneos. Ademais tarefas de pesquisa concorrem com as tarefas administrativas de compartilhamento dos dados pelos pesquisadores e pelas Coordenações dos Programas.

1.2 Justificativa

A literatura inclui algumas opções de ferramentas que podem ser usadas para a autoavaliação de programas. Dentre elas, alguns recursos disponibilizados pela própria CAPES na plataforma Sucupira (CAPES, 2022), bem como ferramentas com recursos de *Business Inteligence*, como o Stela Experta (STELATEK, 2022), dentre outras. Entretanto, as opções da literatura se resumem à exploração de dados gerais dos programas, que podem ser capturados a

partir de repositórios como o lattes ou em sistemas acadêmicos. Isso, em geral, não é suficiente para todas as tomadas de decisões internas dos programas, já que cada um conta com um conjunto próprio de requisitos. Por exemplo, contar a quantidade de publicação de determinado docente, ainda que ponderada pelo peso do qualis, não informa quantos alunos fizeram parte de cada publicação, se havia pesquisador estrangeiro coautorando determinado trabalho, ou qual é porcentagem de docentes de um mesmo PPG que divide a autoria de cada publicação.

Informações internas desta natureza, que contribuem a análises gerenciais relevantes não estão disponíveis em bases gerais, como o lattes. A ausência desse detalhamento interno impede os programas de, por exemplo, avaliar balizas importantes consideradas pela Capes, como internacionalização, quantidade absoluta de produção (a parte absoluta é aquela que considera a quantidade de autores que dividem a produção), participação discente na produção, etc.

Além disso, há ainda de se considerar que cada PPG contempla um conjunto de requisitos próprios a serem avaliados junto à sua área. Por exemplo, o colégio das humanidades pode considerar livros como produção qualificada, enquanto as engenharia poderia considerar patentes, software, etc. Essas particularidades dificultam a programação de ferramentas capazes de absorver cada especificidade. Neste contexto, que a solução proposta se insere, para resolução dos problemas citados, maximizando possibilidades de obtenção de informações e análises.

No mais, a proposição de um modelo de sistema com integração de dados e centralização, em conformidade com as prerrogativas e diretrizes dos órgãos externos de controle viabiliza diferentes visões de análise e emerge como alternativa para embasar a tomada de decisão dos stakeholders (Coordenadores, Gestores de Pesquisa, Pesquisadores) e para ganho de eficiência e consolidação dos processos de controle e auto-avaliação, viabilizando um diagnóstico situacional para melhoria de indicadores de qualidade e atribuindo maior celeridade à recuperação da informação para fornecimento aos órgãos de fomento e controle externos.

1.3 Objetivos

Os principais objetivos do trabalho são apresentados a seguir.

1.3.1 Objetivo Geral

Estabelecer um modelo de integração de dados com especificação e implementação de um sistema de Auto-avaliação para cursos de Pós-Graduação.

1.3.2 Objetivos Específicos

- Analisar os dados necessários à integração e tomada de decisão dos gestores de cursos de Pós-graduação;
- Realizar extração, transformação, carga e integração dos dados para armazenamento centralizado.

- Elaborar pipelines com processos de análise dos dados, seguindo regras e critérios definidos pelo especialista de negócio;
- Desenvolver interface dinâmica de pesquisa, análise e apresentação dos dados, baseados nos critérios de auto-avaliação institucional especificados junto a especialista do domínio.
- Avaliar o modelo de sistema com especialista do domínio.

1.4 Organização do Trabalho

As demais seções descrevem o processo de desenvolvimento deste trabalho. O Capítulo 2 apresenta o referencia teórico norteador, contendo informações que descrevem a área na qual o trabalho se insere, conceitos, arquiteturas e processos de análise, sob a ótica e sustentação de autores com extensa contribuição teórica. No Capítulo 3 a metodologia utilizada é detalhada, apresentando o arcabouço processual, técnico e tecnológico utilizado. No Capítulo 4, os resultados são apresentados após desenvolvimento do produto de software, seu ambiente de aplicação em estudo de caso, com homologação de métodos de análise, e informações obtidas, e por fim a apresentação de trabalhos similares, com a problemática e objetivos que pretendem alcançar. No Capítulo 5, são socializadas as principais conclusões alcançadas, bem como as limitações, pontos fracos e oportunidades de pesquisa e melhoria.

2 REVISÃO DE LITERATURA

2.1 Avaliação da Pós-Graduação

De acordo com MEC (2001), os cursos de Pós-Graduação Stricto Sensu, compreendendo programas de mestrado e doutorado, são sujeitos às exigências de autorização, reconhecimento e renovação de reconhecimento previstas na legislação. A autorização, o reconhecimento e a renovação de reconhecimento de cursos de Pós-Graduação Stricto Sensu são concedidos por prazo determinado, dependendo de parecer favorável da Câmara de Educação Superior do Conselho Nacional de Educação, fundamentado nos resultados da avaliação realizada pela Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES e homologado pelo Ministro de Estado da Educação.

A Avaliação dos Programas de Pós-graduação compreende os processos de Acompanhamento Anual e de Avaliação Quadrienal do desempenho dos programas e cursos que integram o Sistema Nacional de Pós-graduação - SNPG (CAPES, 2020a)

Nos últimos anos o processo de Avaliação de Cursos de Pós-Graduação foi aprimorado, tendo seu foco de análise transferido do acompanhamento da formação docente à aspectos de qualidade das publicações, com base na quantificação de produção e dos canais de publicização do conhecimento científico como periódicos, avaliados pelo Qualis. Além de melhoria dos instrumentos de coleta de dados como formulários (MARQUES; VEIGA A. C, 2020).

O Qualis-Periódicos é um sistema usado para classificar a produção científica dos programas de pós-graduação, no que se refere aos artigos publicados em periódicos científicos. Tal processo foi concebido para atender as necessidades específicas do sistema de avaliação e é baseado nas informações fornecidas por meio do aplicativo Coleta de Dados. Como resultado, disponibiliza uma lista com a classificação dos veículos utilizados pelos Programas de Pós-Graduação para a divulgação da sua produção (CEFET-MG, 2019).

De acordo com Hostins (2006), uma mudança na sistemática de avaliação da Capes foi introduzida, com inserção de indicadores que pudessem discriminar os programas em termos de qualidade e induzir a competitividade, inovação e empreendedorismo também é definido.

Neste contexto, BRASIL (2016) elenca cinco quesitos como premissas para análise: Proposta do Programa; Corpo Docente; Corpo Discente, Teses e Dissertações; Produção Intelectual; e Inserção Social. A partir de ajustes nos pesos e dimensões de cada quesito, que são variáveis por áreas do conhecimento, tal ficha foi usada desde então, inclusive na quadrienal finalizada em 2017.

Após a análise e compilação das notas com base na utilização de pesos definidos pela CAPES. De acordo com Marques e Veiga A. C (2020): Os Programas de Pós Graduação são enquadrados da seguinte forma.

- Cursos com nota 1 e 2 tem funcionamento impedido, pois estas notas são destinadas

aos programas que não alcançam o padrão mínimo de qualidade.

- Aqueles com nota 3 têm o funcionamento autorizado, mas não podem abrir cursos de doutorado;
- os de nota 4 são considerados de bom desempenho, podendo ter o nível 1 de doutorado autorizado.
- A nota 5 é atribuída aos programas com alto nível de desempenho, sendo a nota máxima a programas que tenham apenas mestrado.
- Por fim, as notas 6 e 7 são atribuídas apenas a programas que tenham doutorado, sejam avaliados com alto padrão de excelência e com equivalência a centros internacionais de destaque. Com isso, a ideia da internacionalização, seja da produção como da formação de pesquisadores e docentes, ganhou maior notoriedade.

De modo a atribuir maior transparência ao processo de Avaliação, toda normatização, critérios e dados utilizados são publicizados na plataforma Sucupira (CAPES, 2022). De acordo com Sant'anna (2019), A Plataforma Sucupira ('Sistema Unificado de Currículos e Programas: Identificação de Redes Acadêmicas'), lançada em 2014 pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), disponibiliza dados abertos sobre a avaliação quadrienal dos Programas de Pós-Graduação brasileiros. Estes dados informam a constituição dos corpos docente e discente dos programas, teses e dissertações defendidas, áreas de concentração, projetos e linhas de pesquisa vinculados e sua produção intelectual. A quantidade de descritores dos dados disponíveis na Plataforma evoluiu desde a implantação do Sistema Nacional de Pós-Graduação (SNPG) em 1998 (Capes, 2014). Entre 1997 e 2012, a base de dados referente às teses e dissertações apresentava 28 colunas para cada registro. Desde 2013, este número aumentou para 57 colunas e passou a incluir identificadores numéricos únicos para todos os dados textuais.

O principal recurso que armazena dados sendo utilizado como referência no ambiente cientifico é a Plataforma Lattes (CNPQ, 2022). Segundo Sampaio, Junior e Mena-Chalco (2018) a Plataforma Lattes é um sistema de informação curricular, disponibilizado pelo CNPq, que permite o registro curricular dos pesquisadores brasileiros. Atualmente, a plataforma conta com mais de cinco milhões de currículos cadastrados. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, o Currículo Lattes (CV Lattes) se tornou elemento indispensável e compulsório à análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia. Este trabalho é aplicado ao domínio da Avaliação em Programas de Pós-Graduação, oferecendo uma alternativa, através de um sistema para análise de dados, registro histórico e aprendizado organizacional para instituições interessadas em sua adoção.

2.2 Big Data Analytics

Raghunathan (1999) defende que a qualidade da tomada de decisões melhora quando gerentes possuem um maior conhecimento sobre os relacionamentos entre as variáveis do problema. Nesta perspectiva, a coleta, armazenamento e análise de dados organizacionais

emerge como fator crítico de sucesso para melhoria do processo decisório.

Big Data é um termo que descreve grandes volumes, sob alta velocidade, complexidade e com tipos de dados variáveis que requerem técnicas avançadas e tecnologias capazes de coletar, armazenar, distribuir, gerenciar e analisar as informações (COMMISION, 2012).

Esta definição enumera as três principais propriedades que sistemas dessa categoria devem exibir, de acordo com Gandomi A. Haider (2015):

- Volume: Se refere à magnitude dos dados. De acordo com Marquesone R. F. P (2020), o sistema deve ser capaz de lidar com mudanças repentinas na carga de trabalho sem afetar negativamente o desempenho, ser capaz de suportar conjuntos cada vez maiores de dados, o atendimento à estas premissas pode ser realizado por meio das estratégias de escalabilidade vertical(scale up), ou horizontal(scale out).
- Variedade: se refere à heterogeneidade estrutural em um *dataset*. O avanço tecnológico permitiu o uso de vários tipos de dados (estruturado, semi-estruturado e não-estruturado). Cukier (2010) enuncia que os dados estruturados, representados por estruturas de tabelas em Bancos de Dados relacionais correspondem a apenas 5% de todos os dados existentes. Os dados não estruturados englobam textos, imagens, áudio e vídeos. Já os dados semi-estruturados, embora apresentem algum grau de organização podem ser descritos em diferentes linguagens de descrição, como XML e JSON.
- Velocidade: se refere à taxa na qual dados são gerados e a rapidez na qual deveriam ser analisados. A proliferação de dispositivos digitais como smartfones e sensores tem acelerado a taxas sem precedentes de criação e tem levado a um crescimento da necessidade de análise de dados em tempo real.

Gandomi A. Haider (2015) também cita que outras propriedades tem sido mencionadas e priorizadas na definição de sistemas de *Big Data*, como:

- Veracidade: elencado pela IBM reconhece à necessidade de identificação precisa por meio de dados de fatores que assegurem uma confiabilidade inerente para algumas análises como por exemplo, análise de sentimento em redes sociais.
- Variabilidade e Complexidade: introduzido pela SAS se refere à variação na taxa do fluxo de dados. Muitas vezes, a velocidade não é consistente e possui períodos de pico. Complexidade se refere ao fato que *Big data* são gerados através de uma infinidade de recursos, o que impõe um desafio crítico de conectar, combinar, limpar e transformar os dados recebidos por diversas fontes.
- Valor: atributo introduzido pela Oracle, define que os dados recebidos em sua forma original possuem pouco valor relativo. Entretanto, um alto valor pode ser obtido pela análise de um grande volume destes dados.

Uma vez armazenados e estruturados, os dados podem ser submetidos a incontáveis técnicas de processamento de modo a extrair valor. Chen et al. (2005) elenca uma série de técnicas e tecnologias, como Sistemas de Banco de Dados Relacionais, *DataWarehousing*, ELT, OLAP e BPM. Assim como variados algoritmos, no contexto de Inteligência Artificial capazes

de realizar classificação, clusterização, regressão, associação e análise de redes sociais, C4.5, K-means, Máquina de Vetor Suporte, Apriori, EM, PageRank, AdaBoost, K-vizinhos mais próximos, Naive Bayes e CART.

Considerando o potencial de escalabilidade da aplicação, uma vez que se trata de um processo de negócio dinâmico com produção massiva de dados, que precisam ser estruturados e agrupados, e considerando a multiplicidade de fontes, tal aplicação apresentada neste trabalho, se enquadra no contexto de Big Data, e suas propriedades relacionadas aos V's devem ser objeto de aprimoramento constante de modo a atingir todo potencial possível.

2.3 Consultas Analíticas e DataWarehouse

Segundo Carniel (2021): *Data Warehouse* corresponde a um ambiente que engloba arquiteturas, algoritmos, ferramentas e técnicas analíticas que possibilita a combinação de dados oriundos de diversas fontes (autônomas,, heterogêneas e distribuídas), de modo a torná-las integradas em uma única base de dados. Satisfaz a necessidade de acesso aos dados (ou seja, leitura) sobretudo para subsidiar a tomada de decisões estratégicas.

Aplicações da classe OLAP se beneficiam da implementação de DWs, segundo Microsoft (2022): OLAP é uma tecnologia de banco de dados que foi otimizada para consulta e relatórios, em vez de processar transações. Os dados de origem do OLAP são bancos de dados OLTP (Processamento Transacional Online) que são comumente armazenados em *Data Warehouses*. Os dados OLAP são derivados dos dados históricos e agregados em estruturas que permitem análises sofisticadas. Os dados OLAP também são organizados hierarquicamente e armazenados em cubos em vez de tabelas. É uma tecnologia sofisticada que usa estruturas multidimensionais para fornecer acesso rápido aos dados para análise. Essa organização facilita que um relatório de tabela dinâmica ou um relatório Gráfico Dinâmico para exibir resumos de alto nível, como os totais de vendas em todo o país ou região, e também exibe os detalhes para sites onde as vendas são particularmente fortes ou fracas.

Os critérios de análise definidos e especificados para o sistema são implementados através de consultas analíticas com diferentes níveis de granularidade para atender necessidades específicas de usuários distintos, e para se ter uma idéia de todo contexto em geral, que se insere o sistema. A estrutura de DW é a base para armazenar os dados históricos e para viabilizar as análises temporais.

2.4 Séries Temporais

Previsões são necessárias em muitas situações: decidir se é necessário a construção de novas plantas de energia nos próximos 5 anos, requer previsão sobre demanda futura; Equipes de agendamento em um call center a serem alocadas na próxima semana, dependem da previsão do volume de ligações; estocagem em inventários dependem do estoque necessário; Previsões podem ser necessárias com antecedência de muitos anos(investimentos de capital), ou em

curtíssimo prazo, poucos minutos (roteamento de telecomunicação). Independentemente das circunstâncias envolvidas, previsão é um importante apoio para um planejamento eficiente e efetivo (HYNDMAN R. J. ATHANASOPOULUS, 2018).

De acordo com Chen et al. (2005), previsão de Séries Temporais é um importante campo de pesquisa e área de aplicação. Muito esforço tem sido dedicado nas últimas décadas para desenvolver e melhorar os modelos de predição de séries temporais. Os modelos de séries temporais mais bem estabelecidos incluem: modelos lineares, como média móvel, suavização exponencial e a média móvel integrada autoregressiva (ARIMA); e modelos não lineares, como, modelos baseados em redes neurais e modelos baseados em sistema *fuzzy*. Recentemente uma tendência para combinação de modelos lineares e não lineares para predição de séries temporais tem sido um campo ativo de pesquisa (ZHANG, 2003).

Segundo Souza e Camargo (2004), pode-se definir uma série temporal como sendo um conjunto de dados observados e ordenados segundo parâmetro de tempo e com dependência serial, sendo esse espaço de tempo entre os dados disponíveis equidistantes (horários, diário, semanal, mensal, trimestral, anual, etc).

Mendenhall (1993) define quatro componentes basilares que estão presentes numa representação em série temporal:

A representação linear dos gráficos se enquadra nas Séries Temporais, representando o comportamento dos indicadores ao longo do tempo relacionados à cada programa, pesquisador e a própria instituição. Tais representação capturam aspectos históricos mas facilitam em trabalhos futuros a evolução com utilização de modelos de predição para um comportamento de gestão mais ativo, na correção de rumos e tomada de decisão.

2.5 Business Inteligence

Tronto, Silva e Sant'anna (2011) relata que, os dados que até então eram simples representantes de fatos comuns como nome, endereço, telefone, dentre outros, hoje se sofisticam na representação de imagens, vídeos, sons, dados temporais, indicadores econômicos, planilhas, páginas HTML e estruturas XML, acompanhando as mudanças solicitadas por uma sociedade agora alavancada por outras indústrias, como entretenimento, comunicação e comércio eletrônico.

De acordo com Negash (2006), sistemas de BI combinam dados operacionais com ferramentas analíticas para apresentar informação complexa e competitiva. Desta forma, o BI é uma coleção de tecnologias de apoio à decisão para a empresa destinada a permitir que os trabalhadores do conhecimento, tais como executivos, gerentes e analistas para tomar decisões melhores e mais rápidas. BI é usado para entender os recursos disponíveis na empresa, o estado da arte, tendências e direções futuras nos mercados, as tecnologias e do ambiente regulatório no qual a empresa concorre; as ações dos concorrentes e as implicações das mesmas (CHAUDHURI; DAYAL; NARASAYYA, 2011)

Ainda segundo Chaudhuri, Dayal e Narasayya (2011), a primeira etapa da arquitetura

Quadro 1 – Componentes em Séries Temporais.

Conceitos	**Descrição**
Tendência	As componentes de tendência são frequentemente, aquelas que produzem mudanças graduais em longo prazo. São normalmente provocadas, por exemplo, pelo crescimento constante na população, no produto interno bruto, no efeito da competição, ou por outros fatores que falham na tentativa de produzir mudanças repentinas, mas produzem variações graduais e regulares ao longo do tempo.
Cíclica	As componentes cíclicas são aquelas que provocam oscilações de subida e de queda nas séries, de forma suave e repetitiva, ao longo da componente de tendência.
Sazonalidade	As componentes sazonais em uma série são aquelas oscilações de subida e de queda que sempre ocorrem em um determinado período do ano, do mês, da semana, do dia ou horário. A diferença essencial entre as componentes sazonais e cíclicas e que a primeira possui movimentos facilmente previsíveis, ocorrendo em intervalos regulares de tempo, por exemplo, ano a ano, mês a mês, semana a semana, ou mesmo dia a dia. Já os movimentos cíclicos tendem a ser irregulares, ocorrendo sobre um período de muitos anos
Erro	São tidos como movimentos ascendentes e descendentes da série após a ocorrência de um efeito de tendência, um efeito cíclico, ou de um efeito sazonal. Nas componentes de erro aparecem flutuações de período curto, com deslocamento inexplicável e geralmente são causadas, entre outros motivos, por eventos políticos e oscilações climáticas imprevisíveis.

Fonte: Mendenhall (1993)

de BI seria a definição de qual será a entrada de dados. Após essa seleção é necessário realizar o processo de ETL, juntamente com o processamento dos dados. Após o ETL, tem-se formado o DW ou data mart, com os dados consolidados e armazenados em um mesmo local. Selecionam-se as ferramentas ou técnicas para realizar as análises de BI e, por último, é definida a interface por onde o usuário irá acessar os dados.

Além do suporte a tomada de decisão e gestão de indicadores, o BI pode ser muito eficaz em análises de Big Data. Ao incorporar ferramentas de DM para extrair informações de grandes volumes de dados e disponibilizar esses dados através de uma interface é possível produzir com velocidade insights que podem ser usados como vantagem competitiva no planejamento estratégico das empresas (ARAUJO, 2019).

Os painéis/dashboards disponibilizados na aplicação e seus diferentes níveis de representação e visualização de dados viabilizam a prática de Business Intelligence para as instituições que aderirem, pois seus conjuntos de indicadores podem ser facilmente apresentados sob diferentes modelos de análise.

3 MATERIAIS E MÉTODOS

Foi realizada pesquisa qualitativa aplicada com análise e síntese da problemática cuja proposição da solução foi realizada com apoio de especialista do domínio de gestão educacional em Cursos de Pós Graduação, através de entrevistas para especificação de requisitos.

Além disso foram realizados levantamentos documentais e normativos (legislação aplicada) para compreensão de exigências, restrições e levantamento bibliográfico com revisão de literatura para verificar pesquisas e identificar soluções aplicadas à área e o seu escopo de aplicação.

O ciclo de desenvolvimento da solução consistiu basicamente em especificação, projeto, desenvolvimento e testes, apoiados nas seguintes tecnologias principais:

Quadro 2 – Tecnologias Utilizadas.

Atividade	**Ferramenta**
Modeladem de Negócio / Entendimento do domínio - Representação da visão geral dos processos e expectativas de melhoria com a implentação da solução	Istar
Obtenção dos requisitos de dados e Modelagem de Dados	Dia
Desenvolvimento da Solução	Linguagem de Programação Python 3.8 e Framework Web Django 2.6 com bibliotecas adicionais, conforme Figura x
Análise de Dados	Pandas (MCKINNEY, 2010)
Visualização de Dados	Matplotlib, Pyplot e Seaborn (WASKOM, 2021)

Fonte: Próprio autor

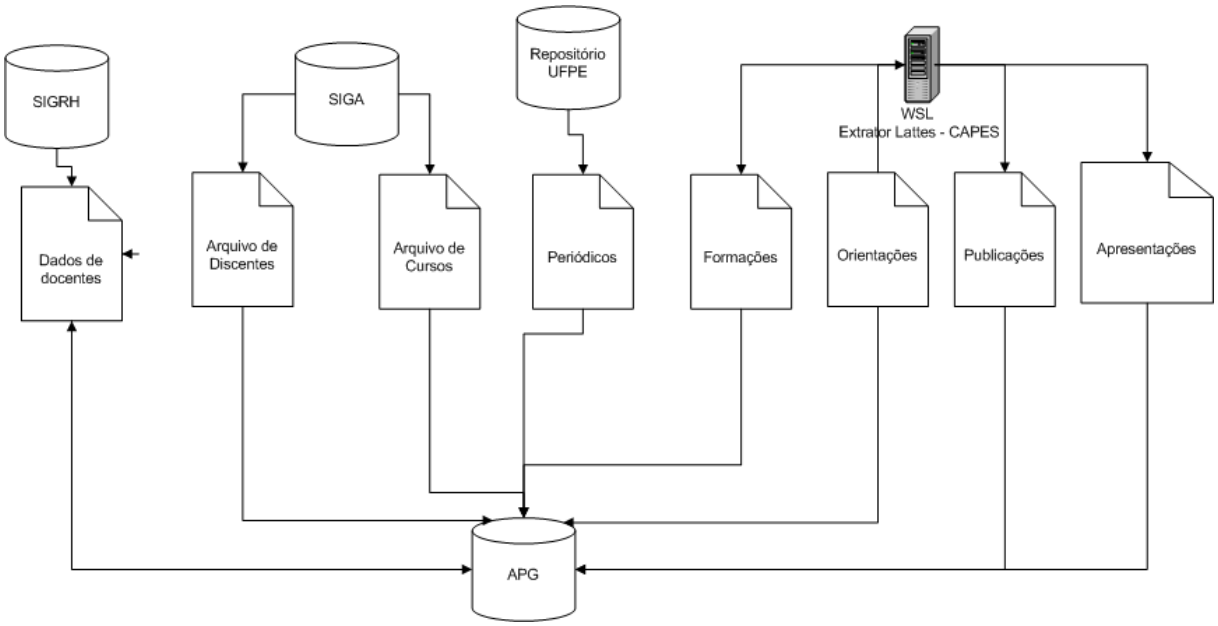
Figura 1 – Stack com tecnologias para desenvolvimento

Fonte: Próprio autor

3.1 Coleta de Dados

A extração de dados para alimentação da solução foi realizada através de importação sobre múltiplas fontes. *Features* de extração foram desenvolvidos seguindo o fluxo descrito na Figura 2:

Figura 2 – Fluxo de dados para alimentação do sistema



Fonte: Próprio autor

- Sistemas institucionais administrativo (SIGRH) e acadêmico (SIGA) forneceram dados para composição da base de usuários relacionadas aos pesquisadores docentes e discentes que foram organizados em grupos, através da vinculação aos programas da instituição. A conexão foi estabelecida com auxílio da biblioteca Sqlalchemy com os drivers nativos dos Bancos de Dados de origem Postgresql e Oracle respectivamente. A extração foi realizada por recuperação direta via banco de dados entre os sistemas, através de consulta SQL.

- O Extrator Lattes foi utilizado para importação dos dados de produção e atividades de pesquisa dos pesquisadores. Os recursos foram obtidos acessando o serviço WSDL-SOAP-REST do CNPQ implementado para viabilizar a publicização dos recursos e para ser consumido por aplicações interessadas, desde que apresentem credenciais de acesso (DMCA, 2022). Utilizando autorização de acesso especial, concedida a instituição para realização das requisições e obtenção dos dados necessários, viabilizou-se a automatização da importação dos dados. Foram extraídos dados relacionados ao conjunto de Formações, Orientações concluídas, Publicações Qualificadas (Artigos, Livros, Capítulos, Softwares, Patentes) e Apresentações de Trabalho em Eventos Científicos. Os dados de entrada foram importados em formato XML através de leitura, mapeamento e parseamento de

dados.

- Do Repositório de dados da UFPE foi implementado um *script* de extração dos principais periódicos, de modo a realizar vinculação de dados para obtenção dos extratos qualis relacionados ao periódico cujas publicações qualificadas foram publicizadas. Isso facilitou o trabalho de classificação das produções realizadas no período, onde estão catalogados 5978 registros de periódicos e seus respectivos identificadores (ISSN's). A base se mostrou insuficiente para identificação de uma quantidade significativa de periódicos para mapeamento do qualis, nessa perspectiva tal fonte necessitará de uma maior intervenção para catalogação, de modo a tornar eficaz a atribuição automática do extrato qualis respectivo.

3.2 Transformação dos Dados

As extrações necessitaram de funções adicionais de transformação e limpeza dos dados, de modo a viabilizar sua organização semântica, realizar os vínculos necessários entre os *datasets*, funções de extração foram distribuídas em cada classe do sistema.

Os dados brutos extraídos das *API's* e dos recursos-fontes foram contextualizados, sumarizados e agrupados para compor os indicadores de gestão. Como a obtenção dos dados relacionados a produções foi em boa parte baseada em *scrapping* de dados, os mesmos precisaram ser padronizados, limpados, estruturados e convertidos em diferentes tipos (*casting*), de modo a viabilizar persistência no banco de dados.

Após extração e armazenamento, uma nova etapa de processamento foi realizada para composição do *datawarehouse* projetado. Este DW foi concebido para armazenar dados com informações contextuais e temporais com os indicadores que foram derivados de dados preexistentes nas tabelas centrais do sistema, aplicando operações de agregação ou de ponderação, neste caso, os pesos foram obtidos junto aos stakeholders sendo aplicados na totalização de alguns dados.

3.3 Validação de Dados

A Validação do modelo foi realizada a partir da exploração da ferramenta por parte de profissionais ligados à autoavaliação de cursos de Pós-Graduação, para os quais a ferramenta foi apresentada.

A partir da apresentação do sistema e suas funcionalidades, foi verificada a sua pertinência na resolução de problemas referentes ao domínio em estudo, possibilitando assim avaliação quanto a utilidade do modelo na consecução dos objetivos propostos influenciando nas dimensões de tomada de decisão e aprendizagem organizacional.

Foram feitos estudos de caso e testes envolvendo um público alvo predefinido e atividades operacionais inerentes às tarefas de autoavaliação, como a análise da produção,

os índices de formação, divisão de artigos por autores e participação discente na produção. Constatou-se que o sistema apresenta um framework inicial com funcionalidades adequadas, corretas e promissoras no sentido de possíveis complementações futuras com atividades mais específicas de cada sistema de autoavaliação.

4 RESULTADOS

Os dados coletados na análise do domínio retornaram o modelo de processo (notação AS IS), apresentado na Figura 4, especificando as atividades realizadas anteriormente à implementação do Sistema. Neste cenário, os dados eram pulverizados em várias fontes de forma independente, sem qualquer integração e dependiam de intervenção dos atores pesquisadores na disponibilização e atualização dos mesmos. Uma planilha era utilizada para armazenar os dados e realizar cálculos importantes para definição dos indicadores de avaliação. O monitoramento era realizado sem qualquer suporte e automatização necessitando de maior esforço para abordagem dos atores, recebimento e análise dos dados. Por fim as planilhas eram analisadas e os dados relevantes eram extraídos para alimentação da Plataforma Sucupira.

Os requisitos funcionais implementados ao sistema, estão descritos na Figura 4. Neste cenário o Sistema de Autoavaliação permite entrada de dados via cadastro, para atualizações simples ou importação de dados integrando diferentes fontes (plataforma lattes, sistemas institucionais e acadêmicos) num banco de dados centralizado. Os gráficos e relatórios serão obtidos do processamento dos dados,a partir de consultas analíticas de granularidade personalizada seguindo os critérios de avaliação especificados. O sistema também é parametrizado para permitir flexibilização nas regras para cálculo dos indicadores.

Estão definidos como usuários principais do sistema:

- Coordenadores de curso: responsáveis por atualizações pontuais de dados, geração de relatórios, consultas e acionamento de importações.
- Pesquisadores discentes: capazes de visualizar informações relacionadas aos seus vínculos
- Pesquisadores docentes: capazes de visualizar informações relacionadas aos seus vínculos.
- Gestor de Pesquisa e Pós Graduação da instituição: responsável pela geração de relatórios analíticos com dados agrupados.

O sistema resultante utiliza um modelo dimensional com dados agrupados e sumarizados, sendo representado na Figura 7, como também obedece a um paradigma relacional de armazenamento no que se refere ao seu núcleo central e cujos dados estão distribuídos de acordo com os requisitos de sistema em 5 módulos:

- Cadastro: Como alternativa é disponibilizado interface para cadastro manual de algumas informações, como forma de agilizar a obtenção de dados pela coordenação e que são utilizados no cotidiano operacional do sistema, entre as quais: usuários, docentes, discentes, coordenação; Inclusão de pesquisadores externos, cujos dados não estão disponíveis nas bases institucionais, mas que podem ser obtidos através da informação do identificador lattes.
- Importação: Provê a interface para obtenção de dados de forma automatizada sob múltiplas fontes. O pipeline de coleta de dados compreende o fluxo de informações necessários para obtenção dos dados para importação e pode ser verificado na Figura 4;

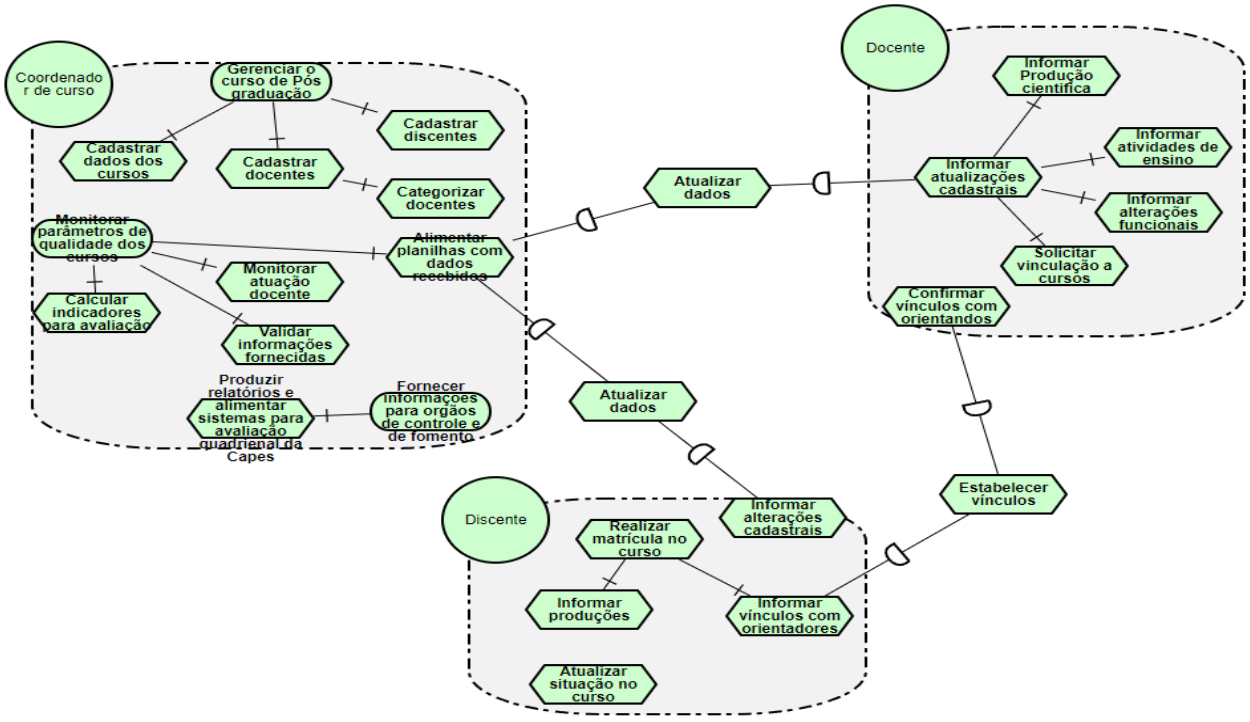


Figura 3 – Modelo As Is

Fonte: Próprio autor

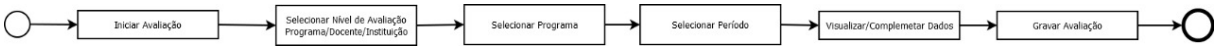
Figura 4 – Pipeline e fluxo de dados para Importação

Fonte: Próprio autor

- Acadêmico: Provê informações relacionadas ao cotidiano acadêmico como vínculos docentes, discentes, matrículas, cargas horárias, orientações, contribuindo na análise de contexto organizacional e acadêmico para tomada de decisão sobre os cursos de Pós Graduação e que são determinantes para consolidação dos indicadores.
- Avaliação: Provê a interface de agrupamento e sumarização dos dados que comporão as consultas analíticas do sistema, as principais estão descritas no Apêndice A. Neste módulo, regras e lógicas são aplicadas para totalização quantitativa de indicadores acadêmicos e de produção, de acordo com o nível de granularidade indicado. O pipeline de avaliação apresenta o fluxo de informações necessárias a realização da avaliação é apresentado na Figura 5.
- Relatórios e Dashboards: Apresenta métodos de visualização de dados e aplicação de filtros sobre os conjuntos de dados derivados, conforme representação no Apêndice C.

Devido à complexidade do domínio, o sistema é organizado em 62 tabelas de dados. Alguns são basilares e compõem fontes para customização da solução e suas interfaces, como definição

Figura 5 – Pipeline e fluxo de dados para Avaliação



Fonte: Próprio autor

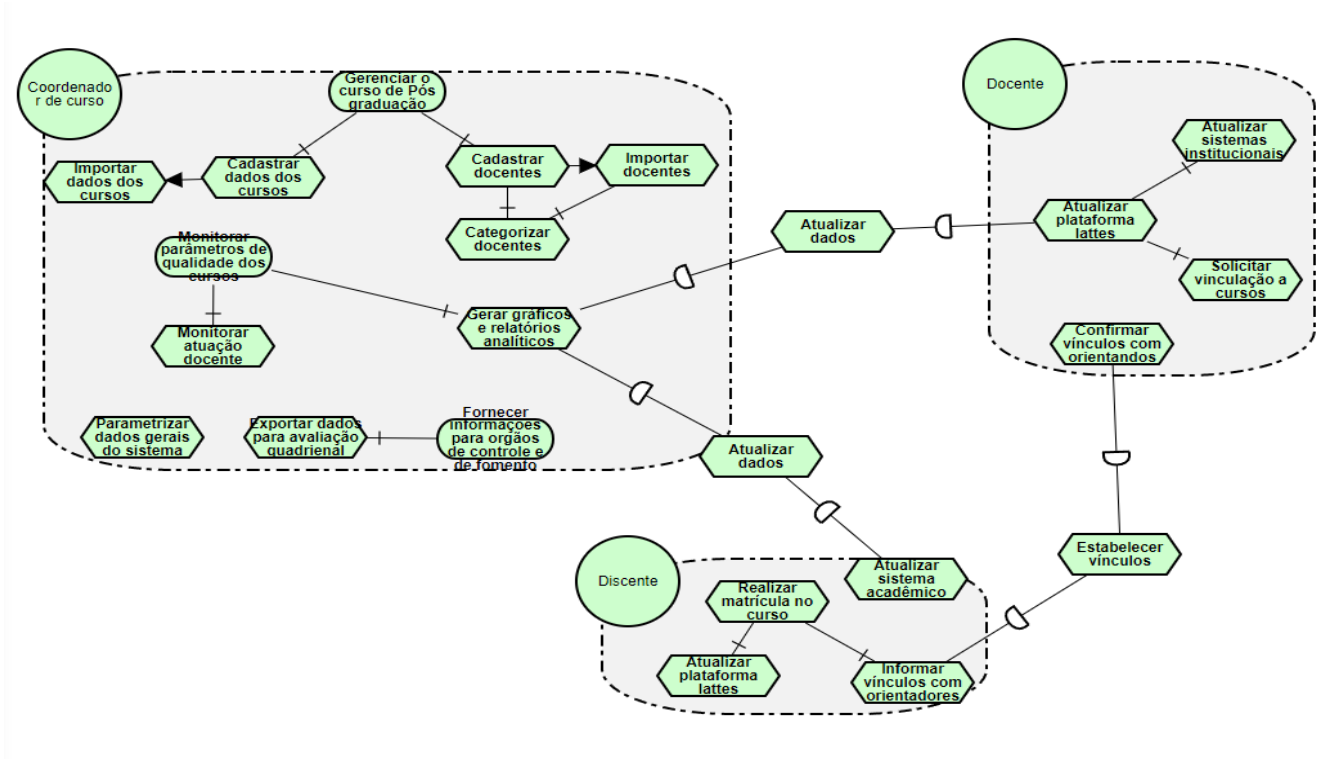


Figura 6 – Modelo to Be

Fonte: Próprio autor

de fontes de extração, tipologia de dados, configurações de dados institucionais, cadastro de conexões externas a banco de dados adicionais, entre muitos outros.

A maioria das tabelas de dados dessa categoria pode ser gerenciado através da interface de administração do sistema (Django admin) (PROJECT, 2022) que provê operações de consulta, atualização, cadastro e remoção. E estarão acessíveis aos administradores da aplicação na instituição e tem seu escopo apresentado na Figura 8.

Conforme representado e comparado nos modelos é esperado uma diminuição na quantidade de processos, enxugamento de tarefas, alimentação centralizada de dados, e uma maior autonomia a equipe de coordenação para obtenção de dados e análise de relatórios, dedicando tempo de análise e tornando mais célere e precisa a tomada de decisão.

4.1 Estudo de Caso

O sistema foi utilizado em ambiente de homologação na Universidade Federal do Vale do São Francisco, com dados reais extraídos das fontes descritas nas seções anteriores. O módulo de Análise de Dados disponível na aplicação apresenta granularidade de análises por pesquisador, programa e instituição.

O *Dataset* que armazena os indicadores de gerenciamento foi construído e armazena 49 atributos, que após importação e processamento, apresenta as características descritas na tabela 1, obtidas via Análise Explorária de Dados, com apoio da biblioteca Pandas.

Ainda sobre análise exploratória de dados relacionadas ao ambiente de aplicação, a Univasf possui atualmente 17 programas de Pós Graduação Stricto-sensu, cujos dados foram extraídos e sumarizados. Os dados sumarizados representaram a composição do corpo de pesquisadores existentes que em conjunto a outros indicadores institucionais coletados após processamento estão especificados na tabela 1:

Tabela 1 – Indicadores da Instituição.

Indicador	Quantitativo
Total de docentes da instituição em exercício	571
Docentes vinculados à Programas de Pós Graduação como Permanente	202
Docentes vinculados à Programas de Pós Graduação como Colaborador	20
Pesquisadores Externos Permanentes	54
Pesquisadores Externos Colaboradores	18
Total de Discentes Matriculados	688
Instituições Nacionais Associadas	20
Instituições Internacionais Associadas	01
Quantidade Total de Publicações	9565
Quantidade Total de Artigos	8999
Quantidade Total de Livros Lançados	248
Quantidade Total de Softwares registrados	79
Quantidade Total de Capítulos de Livros	239
Quantidade de Trabalhos Apresentados em Eventos	11150
Quantidade Total de autores citados em Publicações	27039
Quantidade de Participações de docentes como primeira autoria em publicações	2661
Quantidade de Orientações de Mestrado Concluídas com Orientador	554
Quantidade de Coorientações de Mestrado Concluídas	154
Quantidade de Orientações de Doutorado Concluídas como Orientador	34
Quantidade de Coorientações de Doutorado Concluídas como Orientador	30
Quantidade de Orientações de Discentes do Programa Mestrado	38
Palavras-chaves únicas utilizadas em Publicações	8460

Fonte: Próprio autor

Séries temporais representadas em gráficos de linha foram geradas para representar as variações e tendências nas distribuição de dados. As mesma são utilizadas nos três níveis de

detalhamento no entanto os dados de identificação foram anônimizados para apresentação neste trabalho, considerando a Lei Geral de Proteção de dados mas uma visão geral do dashboard pode ser obtido no apêndice B

Em relação à validação do sistema e das informações obtidas, o stakeholder avaliou como satisfatórios os componentes entregues e definiu como 'de valor agregado' às suas rotinas de análise e monitoramento dos cursos e da produção científica acadêmica com expectativa de implementação em ambiente de produção. De modo a viabilizar reprodutibilidade o código fonte do sistema está disponível em Santos (2022).

4.2 Trabalhos Relacionados

Sampaio, Junior e Mena-Chalco (2018) desenvolveram uma ferramenta para extração de curriculos lattes, implementado com a plataforma R, para utilização no âmbito da FIOCRUZ, extraindo dados em formato XML e realizando estruturação e disponibilização em formato Json. De acordo com os trabalho futuros da publicação, os autores pretendem incorporar mecanismos de análise e ciência de dados para enriquecer a ferramenta.

Abreu e Romao (2017) também idealizaram um extrator de curriculos lattes, com implementação em linguagem C, reunindo dados da Univille, os dados foram armazenados em um *DataWarehouse* e sua análise foi desempenhada com apoio da ferramenta de *Business Inteligence* Tableau, permitindo a construção de alguns gráficos analíticos.

Mena-Chalco J. P. Cesar Junior (2013) implementaram um *script* de extração de curriculum lattes em linguagem Python, denominado Scriptlattes, que utiliza como entrada o id lattes dos pesquisadores realizando a extração e exibição em formato HTML. Tem como componentes um verificador de inconsistência e duplicidade baseado na distância de Levenshtein (NAVARRO, 2001), gráfico de colaboração baseado em geolocalização.

Galego (2013) implementou uma solução baseada em Web Semântica incorporando outras ferramentas entre os seus componentes como Scriptlattes (extração) e Semanticlattes (ontologia e inferência). Esta abordagem permite a exportação dos dados em arquivo OWL e realizou verificação de inconsistência entre dados informados pelo orientador e orientando em relação a formação.

Marques e Veiga A. C (2020) a partir de dados públicos relacionados à avaliação quadrienal 2013 a 2016, construíram um modelo de predição de probabilidade de mudança de nível de classificação de programas de pós graduação, realizando coleta de dados via Web Scrapping e análise com utilização de regressão logística correlacionando as notas aos quesitos com a probabilidade de ascenção ou redução de ranqueamento entre as classes de Avaliação Capes (1 a 7). As variáveis mais significativas apontadas pelo modelo foram: Produção intelectual e do Corpo Discente, Teses e Dissertações.

Há uma variedade de ferramentas disponíveis para extração do curriculo lattes, mitigando o processamento manual dos dados e minimizando a incidência de erros na catalogação, cada uma possui um propósitos e atende a necessidades específicas utilizando abordagens complementares.

Grande parte dos trabalho apontados centraliza o seu foco no processo de extração de dados, tal processo se refere a um dos módulos do sistema especificado neste trabalho trabalho que em conjunto com uma série de componentes visa apoiar o processo de tomada de decisão relacionado à Avaliação de Programas de Pós Graduação. Nesta perspectiva, o escopo não se limita apenas aos dados disponíveis na plataforma lattes, mas à aspectos contextuais acadêmicos e gerenciais que envolvem o contexto da tomada de decisão em cursos deste nível.

A recuperação de dados para a avaliação quadrienal é imprescindível, mas a melhoria do processo em si depende de variáveis adicionais que podem ser incorporadas e são inerentes

às atividades de gestão de ensino e de projetos de pesquisa, requisitos adicionais especificados pelo especialista do domínio permitiram a incorporação de elementos de análise adicionais.

Tal mapeamento e estruturação de dados do domínio oferece subsídios a realização de análises mais complexas e avançadas que deverão ser implementadas em iniciativas futuras, aumentando a escalabilidade das features da aplicação e sua capacidade de adicionar valor aos processos neste contexto.

Ademais o propósito da ferramenta é dar subsídios gerenciais e de apoio à decisão baseada em dados atuais e precisos apoiado em tecnologia open source para composição do portfólio em instituições interessadas, uma vez que, grande maior das implementações utiliza ferramentas de BI proprietárias.

5 CONCLUSÃO

Os processos tendem a ser aprimorados ao longo do tempo, através de redesenho, reestruturação e incorporação de inovações tecnológicas. Tal fato é evidente no processo de Avaliação de Programas de Pós Graduação. Ao mesmo tempo, em que a aplicação de recursos orçamentários vem diminuindo significativamente, o monitoramento da qualidade das publicações e das iniciativas em pesquisa da comunidade acadêmica devem ser reforçados, como forma de viabilizar melhorias na gestão dos programas, tendo em vista que é impossível melhorar, o que não se pode medir e monitorar.

A recuperação da informação para realização das Avaliações previstas pela CAPES para reconhecimento e autorização é uma tarefa custosa, tendo em vista a complexidade de análise de produção, o agrupamento de dados necessários e a disponibilização de mão de obra e tempo docente e da coordenação, que possuem tarefas concorrentes de ensino, pesquisa e extensão.

O sistema proposto visa fomentar a extensão de funcionalidades, integrando dados de vários sistemas para viabilizar uma tomada de decisão mais assertiva, para tal são integradas informações administrativas e acadêmicas oriundas dos sistemas gerenciais internos, associados aos dados extraídos da plataforma lattes, tentando obter respostas que justifiquem a melhoria ou decaimento no desempenho da produção ou atividades acadêmicas dos programas.

De outro modo, também contribui dando subsídios há um monitoramento ostensivo e efetivo, sendo necessário apenas a atualização dos dados, pela atualização das importações, em detrimento de um processo intervalar, apenas voltado ao atendimento de requisitos de obrigatoriedade de informações para composição da avaliação quadrienal.

É esperado uma melhoria também aos processos de acesso, pesquisa e recuperação da informação, na ocasião da submissão de dados para a Plataforma Sucupira, de modo a viabilizar as análises para manutenção de reconhecimento dos cursos.

Outro ponto forte que sistema proposto apresenta, é o ajuste fino de granularidade das informações, cobrindo o monitoramento de avaliação individual dos pesquisadores, com sumarização de dados e geração de gráficos de produção para visualização, como também em relação à avaliação coletiva do programa e avaliação geral da pesquisa a nível de instituição, inferindo comparações quantitativas.

Considerando a abrangência de utilização, o sistema foi projetado de modo a ser totalmente parametrizável, desde as informações do front-end, métodos de importação de dados, até o próprio conjunto de dados que precise ser importado e que se deseje extrair. Desta forma, o custo de manutenção e customização tende a ser baixo.

Com relação a necessidade de dados e alimentação por preenchimento de vários formulários, os usuários envolvidos tendem a induzir um mínimo de esforço possível, uma vez que, a maior parte dos dados é importada, tratada e agrupada, sendo esperado o mínimo de

esforço com preenchimento em formulários ou sistemas. Também espera contribuir à prática, definindo uma cultura de atualização frequente na plataforma lattes pelos pesquisadores, para além da divulgação de ações cientificas, viabilizar a coleta dos indicadores quantitativos necessários para as avaliações.

Á medida que a cultura de gerenciamento baseado em dados for se enraizando na instituição, o sistema define uma plataforma capaz de incorporar funcionalidades no contexto de Ciência de Dados para inferência, classificação e predição. Por ser desenvolvido em Python, a incorporação de bibliotecas nativas da linguagem facilita o desenvolvimento de features dessa natureza com menor esforço de programação e reutilização de classes e algoritmos prontos e validados pela crescente comunidade que trabalha pelo desenvolvimento da linguagem.

Foi possível verificar junto aos stakeholders:

- Que o sistema apresenta licença de utilização livre com modelo de dados, modularizado e parametrizável para instituições interessadas
- Oferece subsídios para uma melhor recuperação da informação para alimentar a plataforma sucupira
- Possibilita aferição automática de indicadores de qualidade para avaliação dos programas

Por fim, Pessoalmente permitiu a implementação de grande parte do ciclo de ciencia de dados, da extração, transformação, carga, a análise, inferência e aplicação de algoritmos de predição e análise

5.1 Limitações

A solução desenvolvida pode contribuir a automação da obtenção de dados sobre pesquisa acadêmica, mas seu melhor desempenho processual depende de obtenção de acesso institucional pelas instituições de pesquisa ao extrator lattes.

Outro ponto, no que se refere aos métodos de obtenção de dados, por utilizar estratégias de consumo baseadas em API's, de acordo com (BIEL, 2016), API é uma solução para conectar componentes de sistemas distribuídos com acoplamento fraco entre si, disponibilizando um conjunto de assinaturas que forneçam serviços e/ou dados. Neste contexto, Mudanças na estrutura dos serviços fornecidos pelos fornecedores de dados, e a disponibilidade destes serviços de terceiros podem influenciar no desempenho da aplicação, durante as sincronizações.

A ausência de padronização nos dados referentes às publicações que são obtidos e descrição de seus autores, minimizam a capacidade de processamento automático pelas regras inicialmente incorporadas no sistema, o que requer intervenção operacional corrigindo dados, antes de acionar o pipeline de avaliação.

Ademais, devido à restrições de tempo, técnicas mais avançadas relacionadas à Ciência de dados não puderam ser aplicadas, no entanto às mesmas possuem previsão futura de implementação e incorporação de *features*.

5.2 Trabalhos Futuros

Empreende-se que a base de dados resultante, viabiliza a aplicação de diferentes análises futuras no contexto de Ciência de Dados, que deverão ser realizadas e incorporadas à aplicação, determinando sua evolução. Como por exemplo, implementação de features que viabilizem a recomendação de pesquisadores com base nos assuntos e linhas de pesquisa que estudam, a partir de regras de associação.

Árvores de decisão podem ser desenvolvidas para análises de desempenho e alocação de recursos orçamentários com base em fatores objetivos para melhoria dos programas e correção de rumos, afim de que avaliações cada vez melhores possam ser alcançadas ao longo do tempo. Como também na definição, com base em critérios objetivos de classificação dos docentes vinculados como Permanentes e Colaboradores, com base na sua produção acumulada.

Como forma de dar publicidade/transparência às iniciativas de pesquisa das instituições, interfaces públicas com sumarização de dados, e métodos de busca e visualização podem ser desenvolvidos, contribuindo às políticas institucionais de Acesso à Informação.

O refinamento de algoritmos, estrutura e arquitetura da aplicação podem ser otimizados e refatorados, como forma de melhorar desempenho, performance e precisão durante as análises ou sincronização de dados.

Além do mais de forma direta é possível elencar como futuras implementações:

- Extensão de funcionalidades em relação às ferramentas existentes;
- Novos modelos para representação dos dados;
- Apoiar o estabelecimento de um processo de avaliação contínuo e institucionalizado com periodização;
- Disponibilização de interface pública (página do curso) / (página do docente com dados sumarizados) de resultados

5.3 Considerações Finais

Por fim, pessoalmente e profissionalmente este trabalho tornou possível a operação de várias etapas do pipeline e do ciclo de vida dos processos de Engenharia/Projeto/Arquitetura para Ciência de Dados, vivenciando na prática atividades relacionadas ao ciclo de desenvolvimento de sistemas no contexto da Ciência de Dados, desde o levantamento de requisitos, passando pela implementação/desenvolvimento e testes até a evolução de um sistema orientado a dados, extração e representação de conhecimento.

Referências

ABREU, L. S.; ROMAO, L. M. Sistema de apoio e gestão de competências aplicado na plataforma lattes. 2017. Citado na página 30.

ARAUJO, G. T. Elaboração de dashboards para análise de big data como vantagem competitiva para o planejamento estratégico de uma organização. **UFOP**, 2019. Citado na página 20.

BIEL, M. Api design. **API University Press**, 2016. Citado na página 33.

BRASIL. Coordenação de aperfeiçoamento de pessoal de nível superior. avaliação quadrienal 2017. 2016. Disponível em: <https://sites.google.com/a/capes.gov.br/avaliacao-quadrienal>. Citado na página 15.

CAPES. Caracterização do sistema de avaliação da pós-graduação. 2020. Disponível em: <shorturl.at/htwx1>. Citado na página 15.

CAPES. Coleta capes: conceitos e orientações. 2020. Citado na página 12.

CAPES. **Plataforma Sucupira**. 2022. <https://sucupira.capes.gov.br/sucupira/public/index.jsf>. Citado 2 vezes nas páginas 12 e 16.

CARNIEL, A. C. **Processamento Analítico de Dados: Introdução e Conceitos**. 2021. <http://moodle.utfpr.edu.br/>. Citado na página 18.

CEFET-MG. Qualis-periódico. 2019. Disponível em: <https://www.bu.cefetmg.br/qualis-periodicos/#:~:text=O%20que%20%C3%A9%3F,artigos%20publicados%20em%20peri%C3%B3dicos%20cient%C3%ADficos.> Citado na página 15.

CHAUDHURI, S.; DAYAL, U.; NARASAYYA, V. An overview of business intelligence technology. **Communications of the ACM**, 2011. Citado na página 19.

CHEN, Y. et al. Time-series forecasting using flexible neural tree model. **Informatin Sciences, v. 174**, 2005. Citado 2 vezes nas páginas 17 e 19.

CNPQ. **Plataforma Lattes**. 2022. <https://lattes.cnpq.br/>. Citado na página 16.

COMMISION, T. F. F. B. data. Demystifying big data: A pratical guide to transforming the business of government. 2012. Citado na página 17.

CUKIER, K. Data, data everywhere: A special report on managing information. 2010. Citado na página 17.

DMCA. **WEB SERVICE DE EXTRAÇÃO CURRÍCULOS Lattes Extrator Manual Extração Currículos**. 2022. <https://usermanual.wiki/Document/Lattes20Extrator2020Manual20Web20Service20de20ExtraC3A7C3A3o20de20CurrC3ADculos.207351363>. Citado na página 22.

GALEGO, E. F. Extração e consulta de informações do curriculo lattes baseada em ontologias. **IME-USP, São Paulo**, 2013. Citado na página 30.

GANDOMI A. HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management, v/35, p. 137-144**, 2015. Citado na página 17.

HOSTINS, R. C. L. Os planos nacionais de pós graduação (pnpg) e suas repercussões na pós graduação brasileira. **Perspectivas, [S. l.] v. 24, n. 1, p. 133-60**, 2006. Citado na página 15.

HYNDMAN R. J. ATHANASOPOULUS, G. **Forecasting principles and practice**. [S.l.: s.n.], 2018. Citado na página 19.

MARQUES, C. A.; VEIGA A. C, B. L. M. Avaliação da pós graduação no brasil: resultados e determinantes da avaliação da capes (2013 a 2016. **Meta: Avaliação, vol. 12, n.37**, 2020. Citado 2 vezes nas páginas 15 e 30.

MARQUESONE R. F. P, P. J. F. Introdução ao big data. **UTFPR**, 2020. Citado na página 17.

MCKINNEY, W. Data structures for statistical computing in python. **Proceedings of the 9th Python in Science Conference**, 2010. Citado na página 21.

MEC. Parecer cne/ces nº 142/2001, aprovado em 31 de janeiro de 2001 - dispõe sobre o funcionamento de cursos de pós-graduação. 2001. Disponível em: <shorturl.at/AJMS7>. Citado na página 15.

MENA-CHALCO J. P. CESAR JUNIOR, R. M. Prospecção de dados acadêmicos do currículo lattes através do scriptlattes. **Bibliometria e Cientometria: reflexões teóricas e interfaces", páginas 109-128. São Carlos: Pedro João Editores**, 2013. Citado na página 30.

MENDENHALL, W. Beginning statistics. **Duxbury Pr**, 1993. Citado 2 vezes nas páginas 19 e 20.

MICROSOFT. **Visão geral do OLAP (Online Analytical Processing)**. 2022. <https://support.microsoft.com/pt-br/office/>. Citado na página 18.

NAVARRO, G. A guide to approximate string matching. **ACM ComputingSurveys, v.33, n.1, p.31-88**, 2001. Citado na página 30.

NEGASH, S. Business intelligence. **Communications of the Association for Information Systems**, 2006. Citado na página 19.

PROJECT, D. **Django Framework**. 2022. <https://www.djangoproject.com/>. Citado na página 27.

RAGHUNATHAN, S. Impacto of information quality and decision-maker quality on decision quality: A theoretical model and simulation analysis. **Decision Support Systems, 26, p 275-286**, 1999. Citado na página 16.

SAMPAIO, R. B.; JUNIOR, A. A. B.; MENA-CHALCO, J. P. e-lattes: Um novo arcabouço em linguagem r para análise do curriculo lattes. **6º EBBC, Rio de Janeiro**, 2018. Citado 2 vezes nas páginas 16 e 30.

SANT'ANNA, H. C. Análise de dados da plataforma sucupira sobre teses e dissertações relacionadas a design da informação (1997-2017). **9th Information Design International Conference**, 2019. Disponível em: <https://sites.google.com/a/capes.gov.br/avaliacao-quadrienal>. Citado na página 16.

SANTOS, L. L. P. **Sistema de AutoAvaliação de Pós Graduação**. 2022. <https://github.com/luamleiverton/apg_project>. Citado na página 29.

SOUZA, R. C.; CAMARGO, M. E. Análise e previsão de séries temporais: os modelos arima. **Rio de Janeiro**, 2004. Citado na página 19.

STELATEK. **Stela Experta**. 2022. <http://site.stelaexperta.com.br/>. Citado na página 12.

TRONTO, A. C. A.; SILVA, J. D. S.; SANT'ANNA, N. Business intelligence: Inteligência nos negócios. 2011. Citado na página 19.

WASKOM, M. L. Seaborn: Statistical data visualization. **Journal of Open Source Software, The Open Journal, v. 6, n. 60**, 2021. Citado na página 21.

ZHANG, G. P. Time series forecasting using a hybrid arima and neural network model. **Neurocomputing, v. 50**, 2003. Citado na página 19.

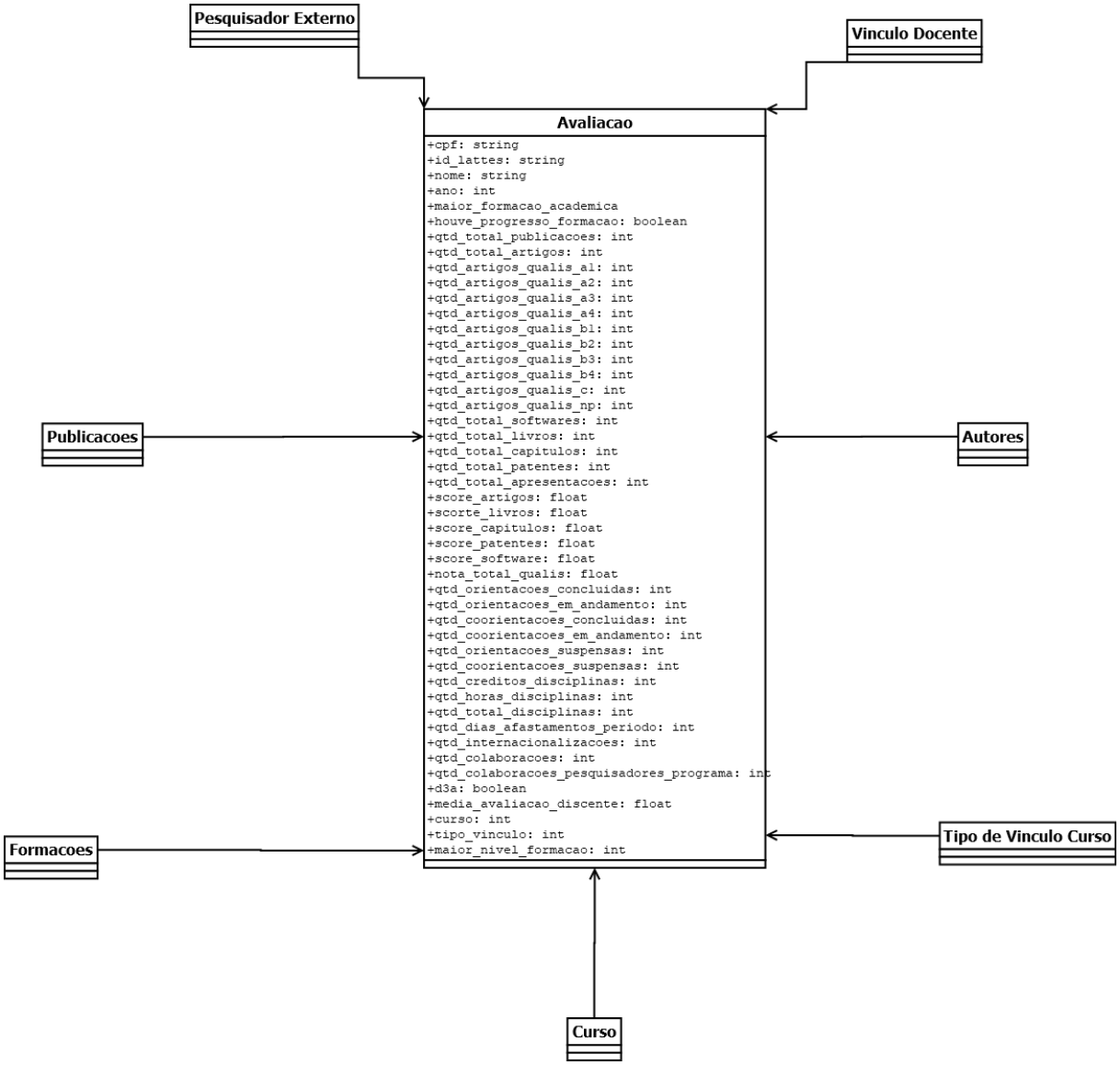
APÊNDICE A – Indicadores do Sistema

Tabela 2 – Indicadores.

Nível Pesquisador 1	Nível Programa	Nível Instituição
Maior Formação no ano de referência	Quantidade de Docentes Permanentes do Programa por ano	
Houve progressão na qualificação	Quantidade de Docentes Colaboradores do Programa por ano	
Quantidade de total produção qualificada no período	Orientações finalizadas no Programa por ano	
Quantidade de artigos por extrato	Quantidade de Orientações em andamento no Programa por ano	
Quantidade de livros publicados	Quantidade de Orientações interrompidas no programa por ano	
Quantidade de capítulos de livros	Quantidade de Publicações Qualificadas no Programa por ano	Conforme Tabela 1
Quantidade de softwares desenvolvidos	Média de Score por ano	
Quantidade de apresentações em eventos científicos	Quantidade de disciplinas ofertadas no Programa no ano	
Quantidade de patentes obtidas	Quantidade de discentes ingressantes por ano	
Score qualis estimado para o Sucupira	Quantidade de discentes	
Quantidade de orientações concluídas	Docentes com maior quantidade de publicações qualificadas	
Quantidade de coorientações concluídas	Docentes com maior score	
Quantidade de orientações em andamento	Média de Avaliação qualitativa do curso	
Quantidade de coorientações em andamento	Quantidade de pesquisadores externos ao curso	
Quantidade de orientações interrompidas	Quantidade de Instituições Associadas	
Quantidade de colaborações com pesquisadores do programa		
Carga Horária de Ensino no Período	Quantidade de docentes com D3A	
Média de Avaliação qualitativa discente no período por Docente		
D3A		
Quantidade de dias afastados por licenças diversas		
Programas de Pós Graduação Vinculados		

APÊNDICE B – Modelo de dados do datawarehouse de referência

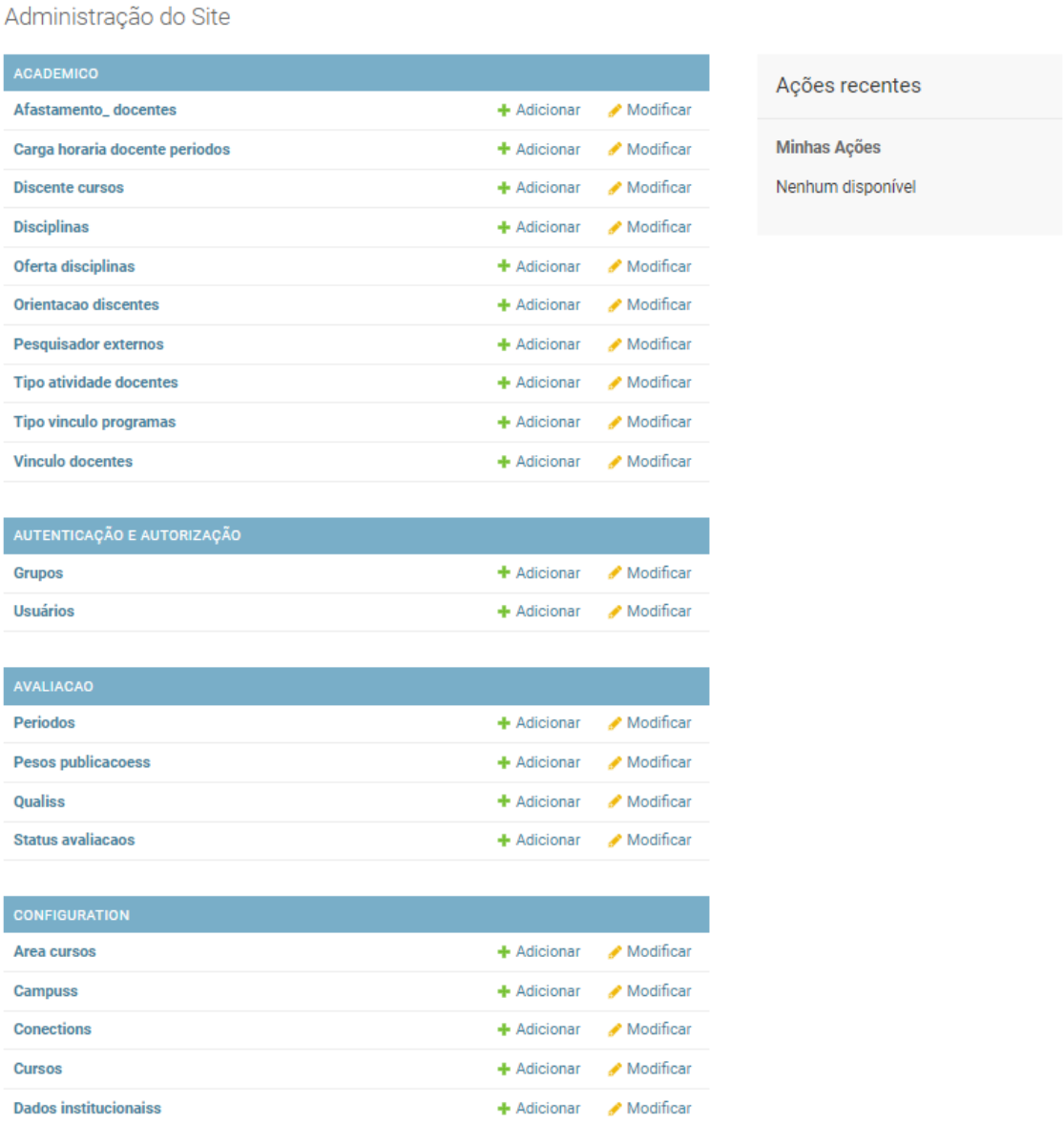
Figura 7 – Modelo dimensional da classe de Avaliação e tabelas relacionadas



Fonte: Próprio autor

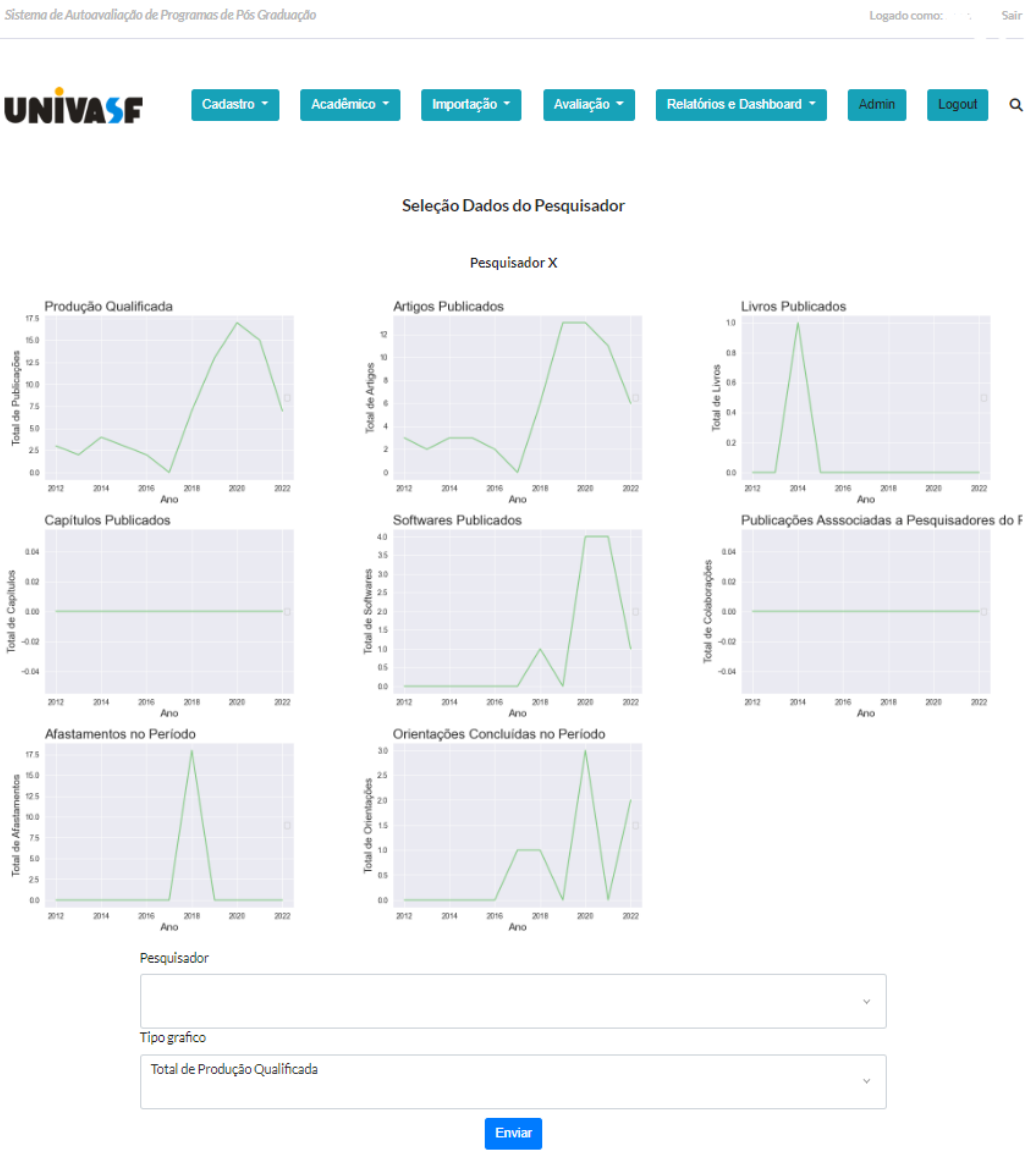
APÊNDICE C – Telas de Exibição do Sistema

Figura 8 – Interface de Cadastro de Dados do Sistema



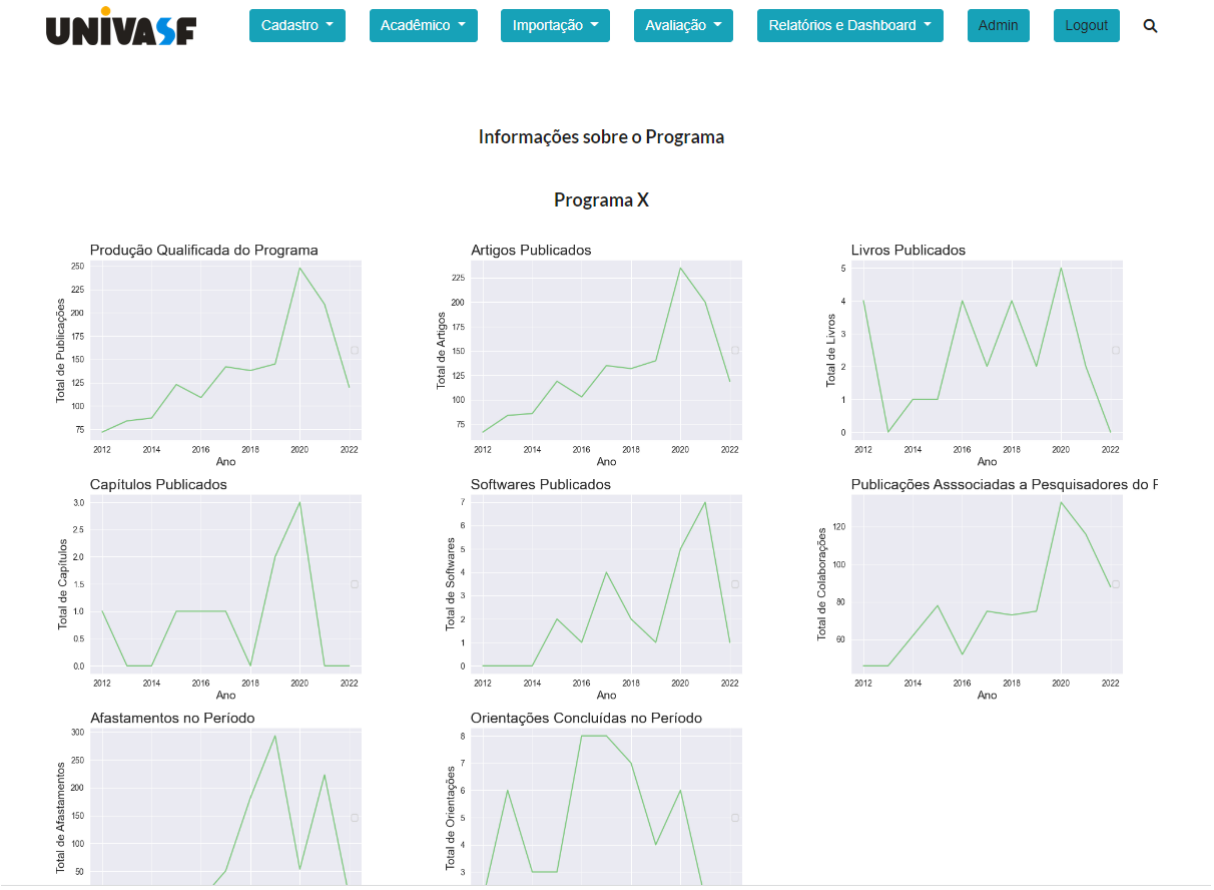
Fonte: Próprio autor

Figura 9 – Telas de exibicação do sistema - Nível Pesquisador



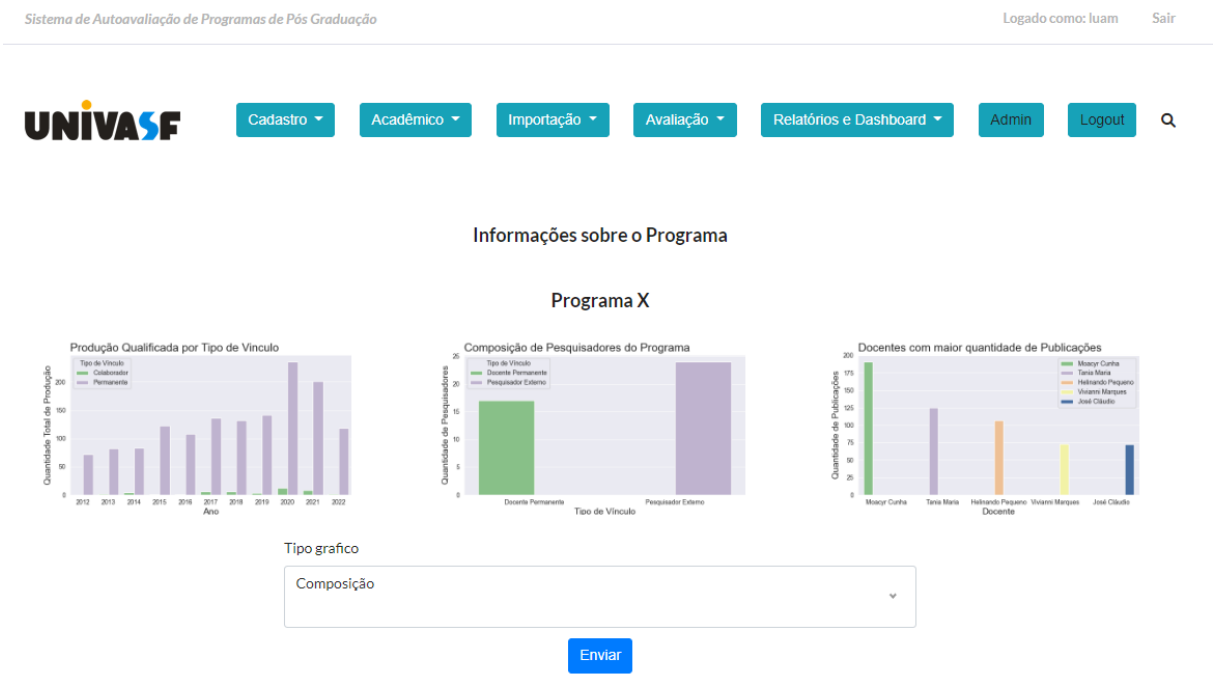
Fonte: Próprio autor

Figura 10 – Telas de exibicação do sistema - Nível Programa 1



Fonte: Próprio autor

Figura 11 – Telas de exibicação do sistema - Nível Programa 2



Fonte: Próprio autor