

# **SCYLAX**

## **Documentação de Entrega de Funcionalidades**

**Anne Magály de P. Canuto <sup>1</sup>, João C. Xavier-Júnior <sup>2</sup> Arthur Costa Gorgônio <sup>1</sup>,  
Cephas A S Barreto <sup>1</sup>, Mateus Firmino Barros <sup>2</sup>, Song J M S Costa <sup>2</sup>**

<sup>1</sup> Departamento de Informática e Matemática Aplicada (DIMAp)  
Universidade Federal do Rio Grande do Norte - Natal, RN - Brasil

<sup>2</sup>Instituto Metr pole Digital (IMD)  
Universidade Federal do Rio Grande do Norte - Natal, RN - Brasil

### **1. Introdu  o**

A pesquisa brasileira tem boa parte de seus dados armazenados na plataforma Lattes (<http://lattes.cnpq.br/>), sistema em que pesquisadores podem armazenar informa  es sobre seus trabalhos cient ficos, art sticos e profissionais.

Tendo em vista a grande quantidade de dados armazenados nessa plataforma, foi iniciado um projeto de pesquisa que tem por objetivo utilizar os dados contidos na referida plataforma, de tal forma que seja poss vel entender o estado atual e tamb m o desenvolvimento da pesquisa brasileira. O principal resultado dessa pesquisa   a plataforma Scylax, um sistema que consome dados da plataforma Lattes, e fornece informa  es relevantes para avalia  o das pesquisas realizadas no Brasil.

Apesar de j  estar em funcionamento, h  uma s rie de desafios relacionados   proposta do Scylax. Esses desafios v o desde a grande quantidade de dados envolvidos em cada an lise ou at  a dificuldade de entender como realizar an lises mais complexas, principalmente as que necessitam de compara  es. Nesse sentido, o Grupo de Pesquisa com Aprendizado de M quina, composto por pesquisadores do Instituto Metr pole Digital (IMD) e do Departamento de Inform tica e Matem tica Aplicada (DIMAp), conduziu pesquisa que objetivou a constru  o de funcionalidades para o Scylax com uso de t cnicas de aprendizado de M quina.

Este documento apresenta, portanto, as funcionalidades constru das pelo grupo mencionado para o Scylax, que s o: an lise de pesquisadores com base em suas pesquisas; an lise da produ  o cient fica de programas de p s-gradua  o; e an lise da produ  o cient fica de redes de colabora  o. Al m da descri  o geral de cada funcionalidade, tamb m s o detalhados os dados utilizados para sua constru  o, as informa  es mostradas, as capturas de tela do prot tipo de sistema e observa  es pertinentes.

## **2. Análise de Produção Científica de Pesquisadores**

### **2.1. Descrição geral**

A análise de produção científica de pesquisador compreende um grupo de informações recuperadas sobre os dados de toda a pesquisa de um dado pesquisador. Para que fosse possível analisar cada pesquisador de forma comparativa, foram utilizadas técnicas não-supervisionadas de Aprendizado de Máquina que, ao final, retornam os dados de forma agrupada. Dessa forma, é possível localizar o pesquisador dentro de um grupo e, ao observar as características gerais do grupo, perceber como está colocado o pesquisador em questão.

### **2.2. Dados utilizados**

Para construção da funcionalidade, foram utilizados os seguintes dados:

- h-index;
- número de colaborações internacionais;
- número de colaboradores internacionais;
- quantidade de publicações em para cada qualis, separadas por década;
- grupo gerado pela técnica de AM.

### **2.3. Funcionalidade exibida no sistema**

A funcionalidade de análise de produção científica de pesquisadores foi dividida em duas partes: na primeira são exibidos recursos relacionados aos grupos e suas métricas gerais; já na segunda é possível escolher um pesquisador e observar algumas de suas métricas.

Para utilizar a funcionalidade, é necessário que o usuário escolha em quantos grupos quer dividir os dados. Essa personalização utiliza como grupos as saídas das técnicas de AM já mencionadas e possibilita a análise dos dados de forma mais ou menos segregada, de acordo com a necessidade do usuário. Após a escolha do número de grupos, são exibidas informações gerais sobre as publicações de cada grupo, inclusive de forma temporal. Na segunda parte, como já mencionado, é possível digitar o nome de um pesquisador e visualizar informações de sua pesquisa, também incluindo uma avaliação gráfica que considera o tempo.

#### **2.3.1. Análise dos grupos - informações exibidas**

- a. número de publicações por década para cada grupo;
- b. h-index médio para cada grupo;
- c. número médio de colaborações internacionais para cada grupo;
- d. média de pesquisadores em colaborações internacionais para cada grupo;
- e. número de publicações por qualis em cada década ([1] 1981 a 1990, [2] 1991 a 2000, [3] 2001 a 2010, e [4] 2011 a 2020) por grupo.

### 2.3.2. Análise dos grupos - capturas de tela

A Figura 1 apresenta um gráfico referente às publicações por década para cada grupo. No eixo horizontal temos os grupos e no eixo vertical a quantidade de publicações. Cada grupo apresenta quatro barras verticais e cada barra representa uma década. O gráfico sempre vem acompanhado de algumas dicas, apresentadas ao clicar no botão com o símbolo interrogação.

A Figura 2 contém 3 gráficos. O primeiro nos informa o H-index médio por grupo. Neste, os grupos estão dispostos no eixo horizontal e o H-index médio no eixo vertical. O segundo diz respeito ao número médio de colaborações internacionais de cada grupo. Neste, os grupos estão dispostos no eixo horizontal e a média de colaborações no eixo vertical. O terceiro gráfico apresenta a média de pesquisadores em colaborações internacionais para cada grupo. Neste, os grupos estão dispostos no eixo horizontal e a média de colaboradores no eixo vertical.

A Figura 3 e a Figura 4 contêm gráficos referentes às publicações por grupo e por qualis em quatro décadas (1981 a 1990; 1991 a 2000; 2001 a 2010; e 2011 a 2020). No eixo horizontal de cada gráfico temos grupos e no vertical a quantidade de publicações. Cada grupo tem nove barras e cada barra representa um qualis conforme a legenda no canto superior direito de cada gráfico.

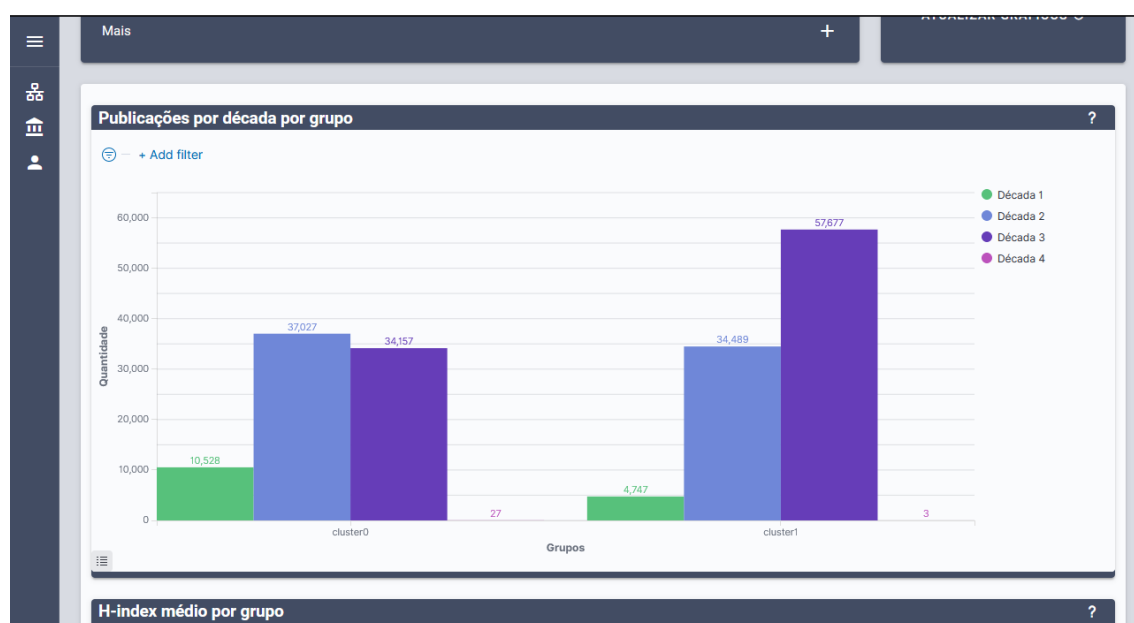
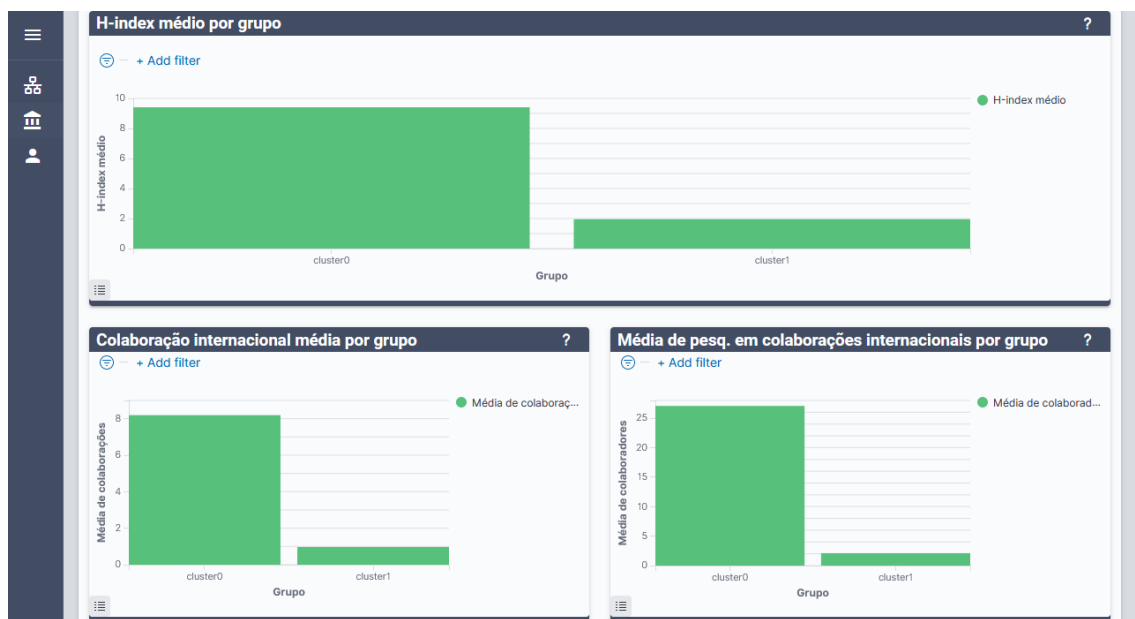
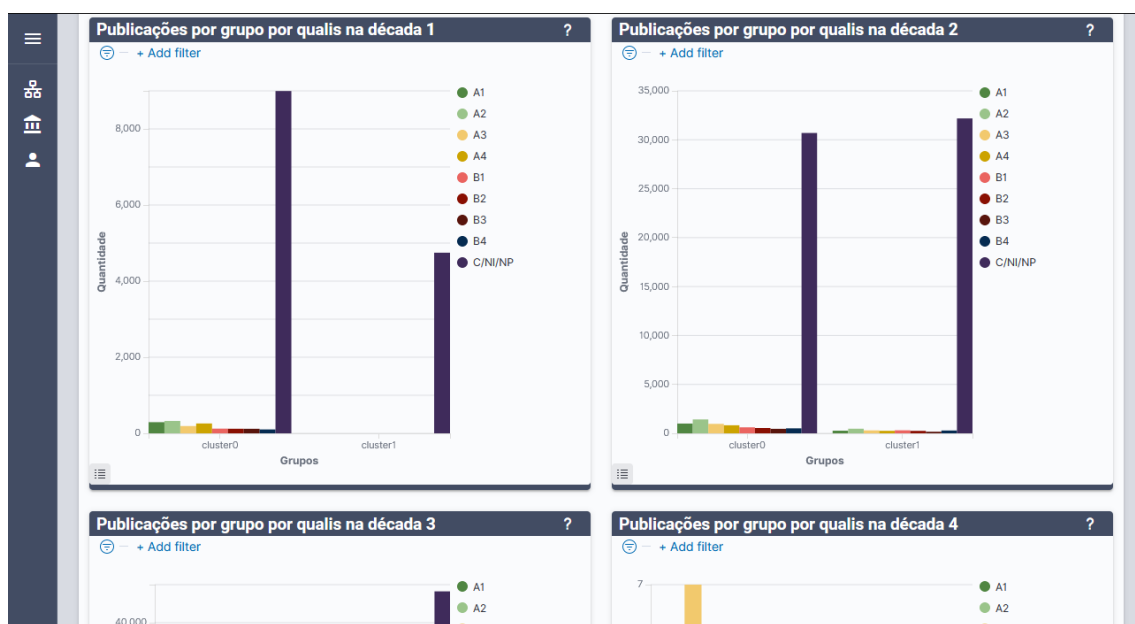


Figura 1 – Gráfico do número de publicações por década para cada grupo



**Figura 2 – Gráficos do h-index médio, colaborações internacionais médias e média de pesquisadores em colaborações internacionais por grupo**



**Figura 3 – Publicações por década e qualis (décadas [1] e [2])**



**Figura 4 – Publicações por década e qualis (décadas [3] e [4])**

### 2.3.3. Análise do pesquisador - informações exibidas

- h-index do pesquisador;
- número de publicações por década para o pesquisador;
- número de colaborações internacionais do pesquisador;
- número de colaboradores internacionais que publicaram com o pesquisador;
- número de publicações do pesquisador, separadas por qualis, em cada década ([1] 1981 a 1990, [2] 1991 a 2000, [3] 2001 a 2010, e [4] 2011 a 2020).

### 2.3.4. Análise do pesquisador - capturas de tela

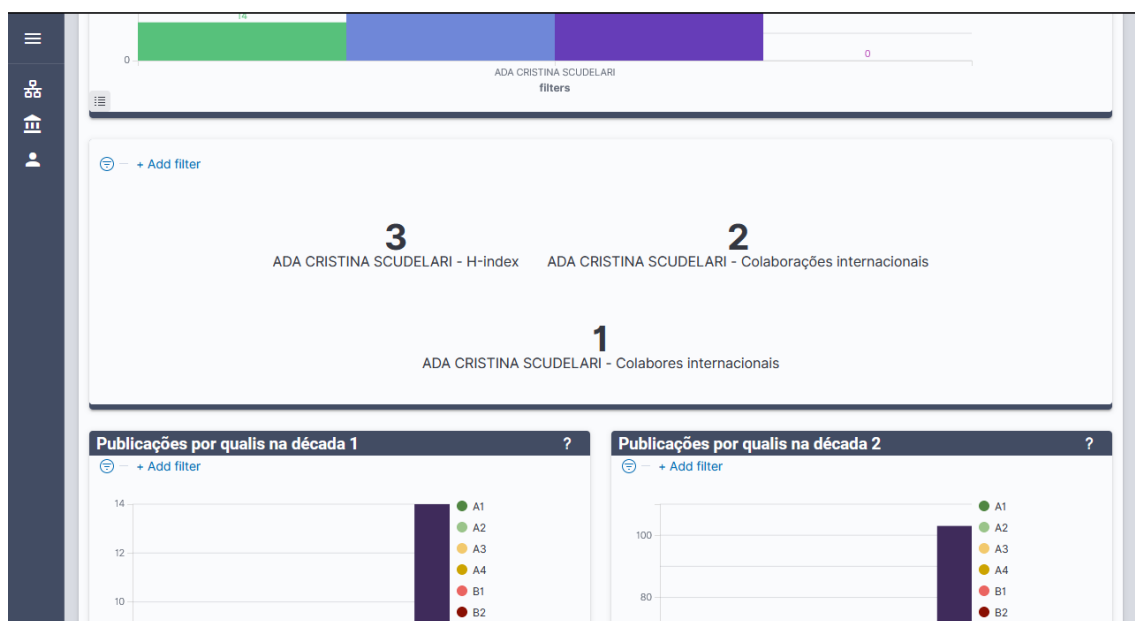
A Figura 5 mostra a área de busca para um pesquisador. Essa figura também contém o gráfico das publicações por década do pesquisador escolhido, com o eixo horizontal apresentando o pesquisador e o vertical a quantidade de publicações. O pesquisador contém 4 barras e cada barra representa uma década conforme a legenda no canto superior direito.

A Figura 6 contém algumas métricas do pesquisador. São exibidos o h-index, número de colaborações internacionais e o número de colaboradores internacionais relacionados ao pesquisador analisado.

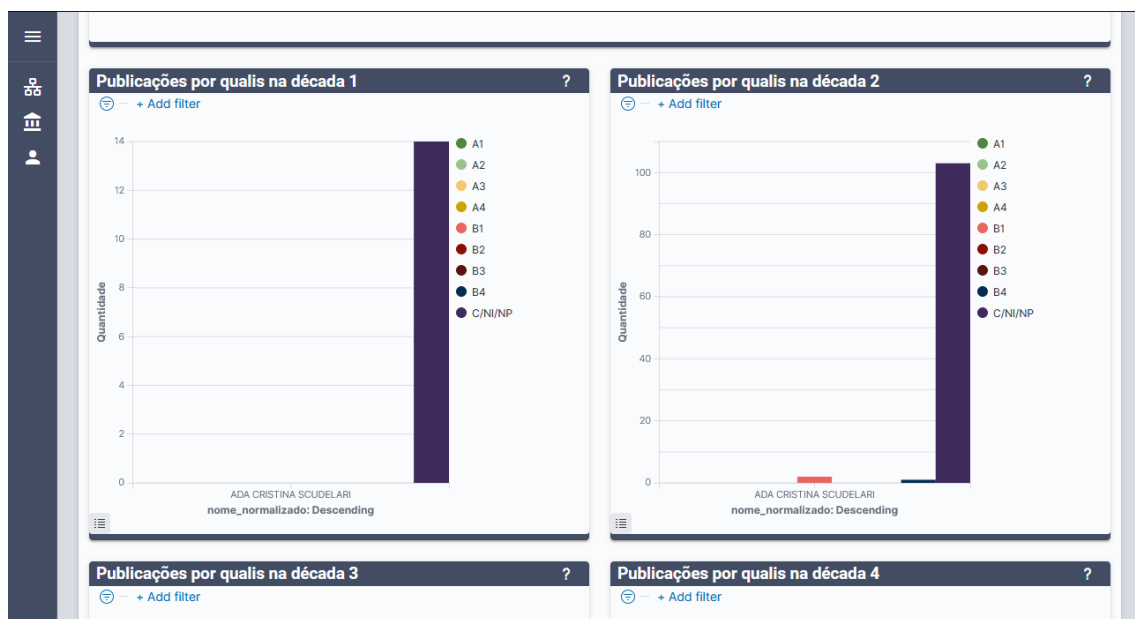
A Figura 7 e a Figura 8 mostram a quantidade de publicações por qualis em cada década para o pesquisador escolhido. No eixo horizontal temos o pesquisador escolhido e no vertical a quantidade de publicações. No gráfico de cada década temos nove barras em que cada uma representa a quantidade de publicações com determinado qualis.



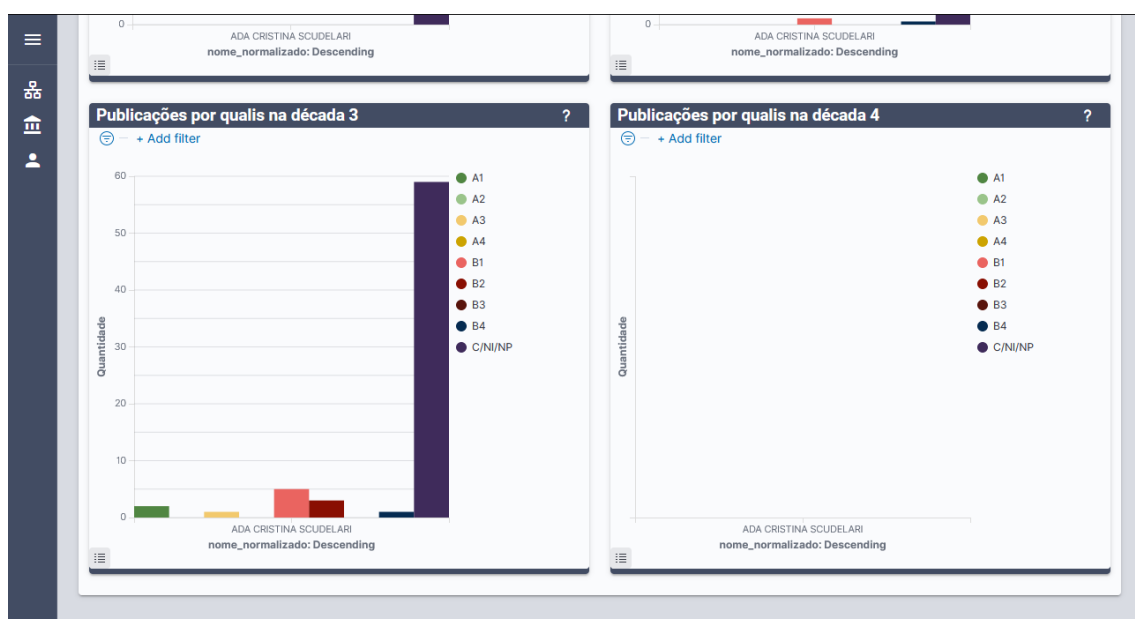
**Figura 5 – Publicações por década de um pesquisador**



**Figura 6 – H-index, colaborações internacionais do pesquisador e colaboradores internacionais nos trabalhos do pesquisador**



**Figura 7 – Total de publicações do pesquisador por qualis (décadas [1] e [2])**



**Figura 8 – Total de publicações do pesquisador por qualis (décadas [3] e [4])**

### **3. Análise de Produção Científica de Programas de Pós-graduação**

#### **3.1. Descrição geral**

De forma análoga à produção científica de um pesquisador, a produção de um programa de pós-graduação é composta por todas as publicações vinculadas a autores que pertencem a certo programa de pós-graduação. A base de dados utilizada para esta análise foi construída utilizando os dados de cada pesquisa científica, seus autores, métricas relacionadas e os programas aos quais a pesquisa em questão está associada. Sobre esses dados foram executadas técnicas de agrupamento, e então, utilizados os grupos gerados como fonte de informação para a construção dos gráficos.

#### **3.2. Dados utilizados**

Para construção da funcionalidade, foram utilizados os seguintes dados referentes a cada produção:

- quantidade de autores;
- lista de autores;
- ano da produção;
- tipo da produção;
- citescore;
- quartil citescore;
- percentil;
- quartil percentil;
- qualis;
- quantidade de programas de pós-graduação envolvidos;
- lista de programas de pós-graduação envolvidos.

#### **3.3. Funcionalidade exibida no sistema**

Esta funcionalidade foi desenvolvida com o objetivo de analisar as produções científicas de um programa de pós-graduação. De forma similar à funcionalidade de análise de produção científica de pesquisadores, é necessário selecionar a “quantidade de grupos” (divisões) pela qual serão divididos os dados. Com o número de grupos escolhido, há a análise geral dos índices de cada grupo e também a possibilidade de comparar um programa alvo com outro(s) programa(s).

Ao gerar os gráficos, também há a divisão em duas partes: na primeira, é mostrada a análise dos grupos como um todo; já na segunda, são exibidas as informações da comparação entre um programa alvo e outros programas, todos escolhidos pelo usuário. A seguir, são apresentadas as métricas utilizadas para cada parte da funcionalidade bem como as capturas de tela do protótipo de sistema desenvolvido.



### **3.3.1. Análise dos grupos - informações exibidas**

- a. quantidade de programas por grupo;
- b. quantidade de publicações por ano, em cada grupo;
- c. citescore;
- d. quartil citescore;
- e. percentil médio;
- f. quartil percentil médio;
- g. quantidade de publicações separados por qualis;
- h. quantidade de publicações separados pelo ranking.

### **3.3.2. Análise dos grupos - capturas de tela**

A Figura 9 exibe o início da área de análise da produção científica de programas de pós-graduação. Nesta área se escolhe o número de grupos a serem analisados. A figura também mostra o gráfico da quantidade de programas por grupo. No eixo horizontal temos os grupos e no vertical a quantidade de programas.

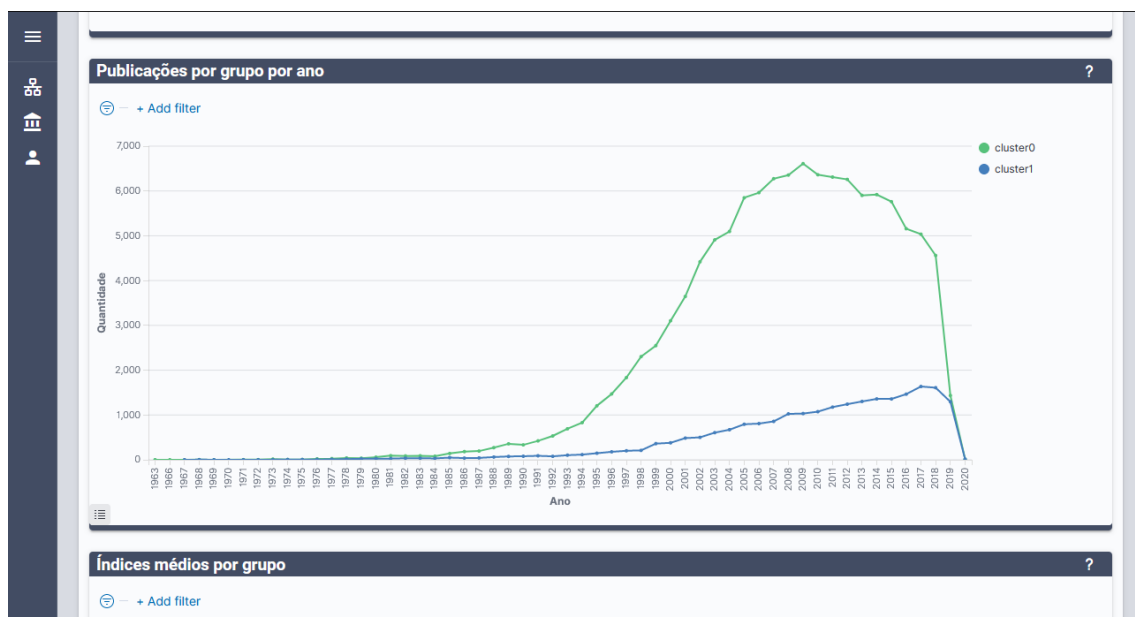
A Figura 10 exibe o gráfico das publicações por ano para o número de grupos escolhido. Temos no eixo horizontal o ano e no eixo vertical o número de publicações. Cada linha representa um grupo conforme as cores exibidas na legenda que está no canto superior direito.

A Figura 11 mostra a média de 4 índices para cada grupo. No eixo horizontal temos os grupos e no vertical a média de cada índice. Cada grupo tem 4 barras, uma para a média de cada índice. Observando a legenda no canto superior esquerdo podemos saber qual barra representa qual índice.

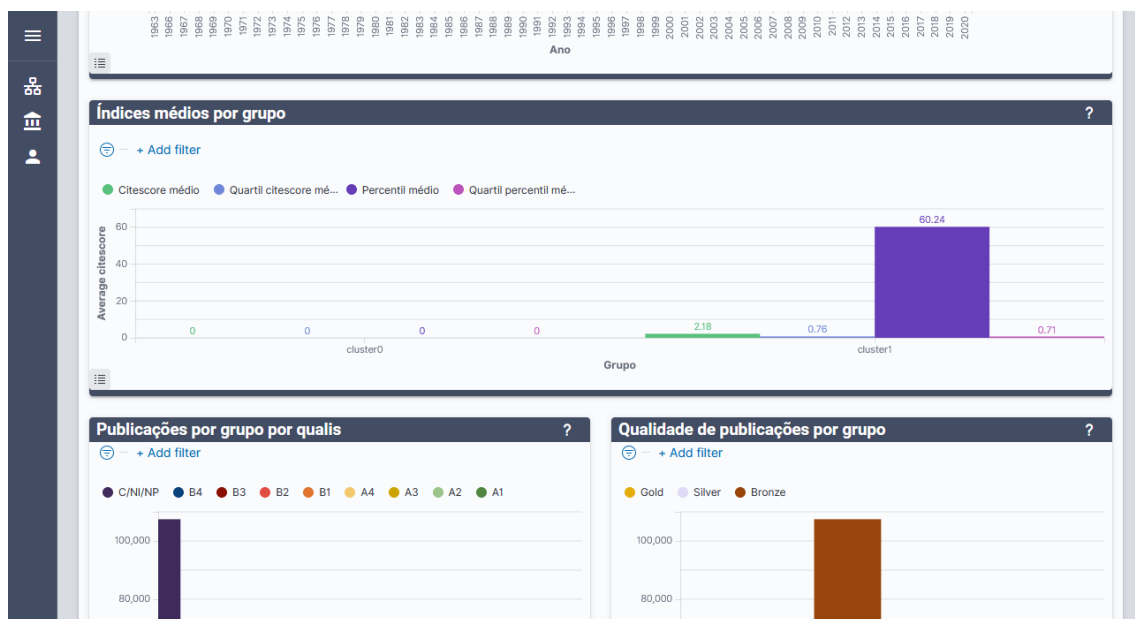
A Figura 12 exibe dois gráficos. O primeiro gráfico diz respeito à quantidade de publicações por qualis para cada grupo. Neste gráfico podemos observar os grupos no eixo horizontal e a quantidade de publicações no eixo vertical. Cada grupo tem 9 barras, uma para cada qualis. O segundo gráfico mostra a quantidade de publicações para cada grupo utilizando uma métrica “ouro, prata e bronze”. De forma análoga ao primeiro gráfico, temos os grupos no eixo horizontal e o número de publicações no eixo vertical. Cada grupo aqui possui 3 barras, cujos significados podem ser consultados na legenda.



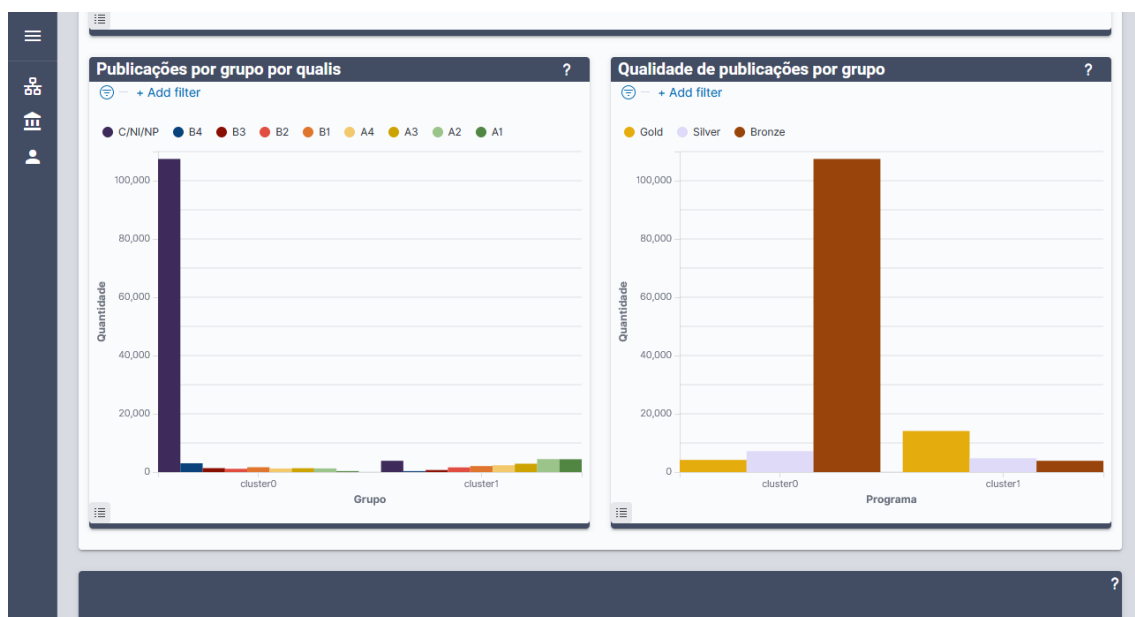
**Figura 9 – Seleção da quantidade de grupos e gráfico quantidade de produções por grupo**



**Figura 10 – Gráfico da quantidade de publicações para cada um dos grupos por ano**



**Figura 11 – Gráfico dos índices citescore, quartil citescore, percentil médio, quartil percentil médio por grupo**



**Figura 12 – Gráficos da quantidade das publicações por qualis e por ranking (ouro, prata, bronze)**

### 3.3.3. Análise do programa de pós-graduação frente a outros - informações exibidas

- citescore médio;
- quartil citescore médio;
- percentil médio;
- quartil percentil médio;

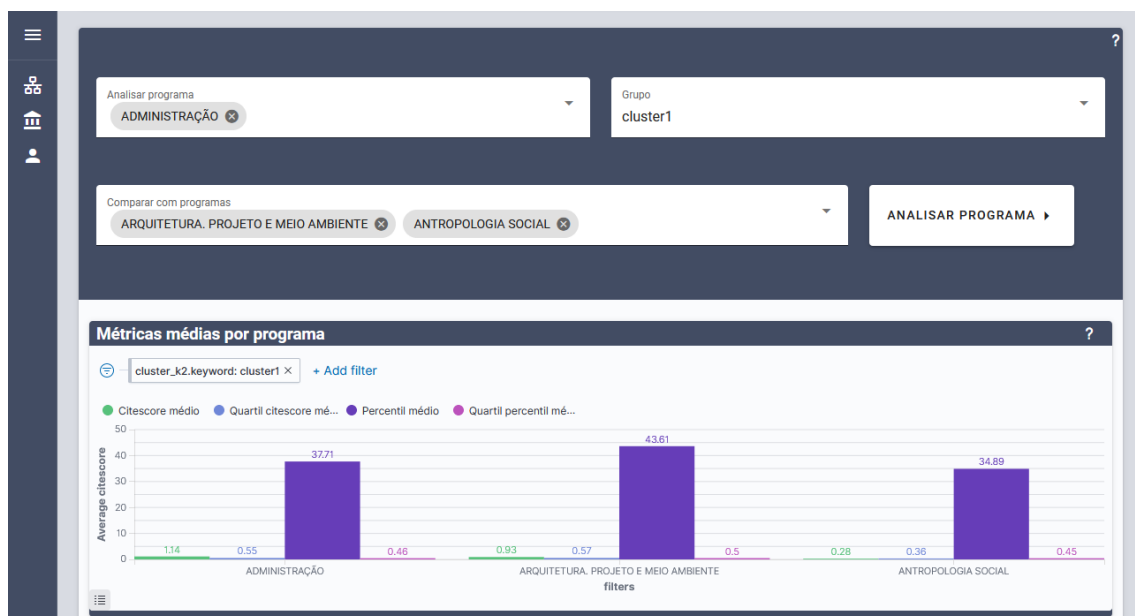
- e. quantidade de publicações por qualis;
- f. quantidade de publicações por ranking;
- g. distribuição das publicações dos programas em relação ao qualis.

#### **3.3.4. Análise do programa de pós-graduação frente a outros - capturas de tela**

A Figura 13 exibe a área do sistema em que podemos analisar um programa de pós-graduação escolhido. Nesta área podemos escolher se estaremos analisando o programa dentro de um grupo ou independente disto. Podemos também realizar uma análise comparativa, informando o nome de alguns programas de pós-graduação na área “comparar com programas”. Ainda nesta figura podemos observar um gráfico com a média de algumas métricas para os programas que escolhemos comparar. No eixo horizontal deste gráfico temos os programas e no eixo vertical o valor médio das métricas. Cada programa tem 4 barras, e cada barra representa uma métrica diferente.

A Figura 14 possui dois gráficos. O primeiro nos traz o número de publicações por qualis dos programas escolhidos para comparação. No eixo horizontal temos os programas de pós-graduação e no eixo vertical temos a quantidade de publicações por qualis. Cada programa possui barras correspondentes a cada qualis, a relação entre a barra e o qualis que ela representa pode ser vista na legenda que está no canto superior esquerdo. O segundo gráfico nos traz a quantidade de publicações para cada programa escolhido utilizando uma segmentação qualitativa “ouro, prata e bronze”. Temos aqui, no eixo horizontal os programas de pós-graduação e no eixo vertical a quantidade de publicações. Cada programa tem 3 barras cujos significados podem ser consultados na legenda que está no canto superior esquerdo.

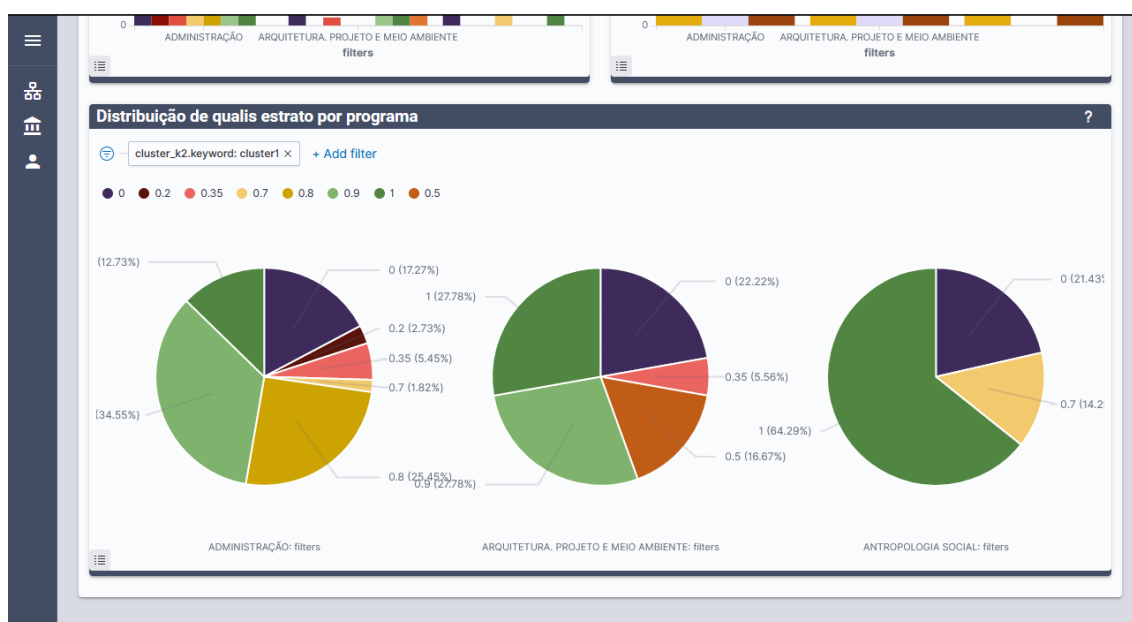
A Figura 15 traz um gráfico de pizza para cada programa escolhido. O gráfico contém a distribuição de qualis estrato para cada programa. O significado de cada fatia do gráfico pode ser consultado na legenda que se encontra no canto superior esquerdo do gráfico.



**Figura 13 – Parte da funcionalidade com a seleção dos programas a serem analisados e gráficos com os índices dos programas escolhidos**



**Figura 14 – Quantidade de produções por qualis e de produções por ranking (ouro, prata e bronze) dos programas selecionados**



**Figura 15 – Gráfico de pizza de cada programa referente a distribuição do qualis das publicações**

## 4. Análise de Redes de Colaboração

### 4.1. Descrição geral

Entende-se, para fins da pesquisa realizada, que uma rede de colaboração é dada por um grupo de autores que possui pelo menos uma publicação em comum. Além desse conceito, também é utilizado o conceito de sub-rede, que compreende qualquer subconjunto entre os autores que compõem uma rede de colaboração.

Esta funcionalidade objetiva analisar qualitativamente uma rede de colaboração e também identificar seus pontos mais fortes e menos fortes. Isso é feito através dos índices dos autores que compõem uma rede e também através dos índices das publicações produzidas pela rede.

No sistema, a funcionalidade foi dividida em três partes: resumo - produção; resumo - colaboração; e detalhes. Na primeira, resumo - produção, são exibidos os dados gerais da rede como um todo. Na segunda, resumo - colaboração, são exibidas informações relacionadas à colaborações internacionais com participação da rede. Na terceira e última parte, detalhes, são detalhadas informações relativas às sub-redes possíveis para a rede de colaboração em questão.

### 4.2. Dados utilizados

Para a construção da funcionalidade, foi criada um *dataset* com foco nos pesquisadores e com os seguintes dados:

- o pesquisador é docente?;
- tipo da produção;
- subtipo da produção;
- número de produções do pesquisador;
- lista de produções do pesquisador;
- código da área do conhecimento;
- nome do programa de pós-graduação do pesquisador;
- h-index do pesquisador.

Com base na identificação dos autores que compõem uma certa rede, o *dataset* mencionado serve de primeira fonte para consulta e construção de uma base de dados dinâmica, referente à rede, e que contém os seguintes dados:

- lista de trabalhos realizados pela rede;
- índices dos autores;
- índices das produções da rede;
- mesmos dados para todas as sub-redes.

### **4.3. Funcionalidade exibida no sistema**

Nessa funcionalidade, é necessário que o usuário informe quais pesquisadores serão analisados como uma rede de colaboração. Após indicar os pesquisadores, os trabalhos científicos em comum são recuperados dos dados e então são apresentados os gráficos.

Como já citado, as três partes da funcionalidade foram pensadas para fornecer um resumo da produção feita realizada pela rede (resumo - produção), dar uma visão geral sobre as colaborações internacionais realizadas pela rede (resumo - colaboração) e fornecer detalhes da rede, inclusive mostrando aspectos relacionados às sub-redes (detalhes).

A seguir serão apresentadas, de acordo com a divisão da funcionalidade, as informações mostradas e as capturas de tela do sistema protótipo desenvolvido.

#### **4.3.1. Análise da rede de colaboração (resumo - produção) - informações exibidas**

- a. total de publicações da rede;
- b. intervalo de atividade;
- c. citescore médio dos trabalhos produzidos pela rede;
- d. número de publicações por tipo;
- e. número de publicações por qualis.

#### **4.3.2. Análise da rede de colaboração (resumo - produção) - capturas de tela**

A Figura 16 exibe a parte do sistema em que podemos selecionar pesquisadores para compor uma rede de colaboração. Para tal basta digitar o nome dos pesquisadores e adicionar

cada um à pesquisa. No campo rede de colaboração podemos ver os pesquisadores que integram a rede escolhida.

A Figura 17 traz o total de publicações da rede, a média de publicações por ano, o intervalo de atividade e o citescore médio da rede de colaboração.

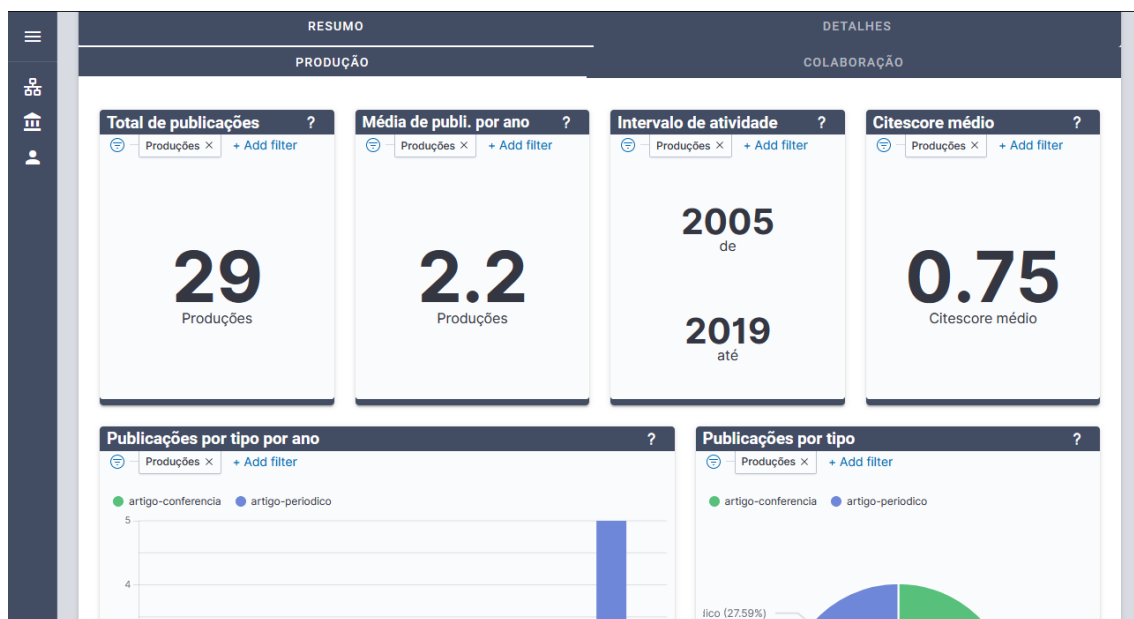
A Figura 18 mostra dois gráficos. O primeiro diz respeito ao número de publicações por tipo por ano. No eixo horizontal deste gráfico temos os anos e no eixo vertical a quantidade de publicações. As barras deste gráfico tem duas cores, azul representa periódicos e verde representa conferências. O segundo gráfico é um gráfico de pizza que mostra as produções por tipo para a rede de colaboração em questão. A fatia verde representa conferências e a fatia azul representa periódicos.

A Figura 19 também possui 2 gráficos. O primeiro é um gráfico de pizza que nos traz as publicações por qualis para a rede de colaboração escolhida. Cada fatia representa um qualis e o tamanho da fatia a porcentagem de publicações com o respectivo qualis. Para saber qual fatia representa qual qualis temos uma legenda no canto superior esquerdo. O segundo gráfico traz informações relativas às produções por qualis por ano para a rede de colaboração em questão. No eixo horizontal temos os anos e no eixo vertical a quantidade de publicações. Cada ano está representado por uma barra e cada barra tem subdivisões com cores diferentes para cada qualis. A legenda indica a cor de cada barra para os qualis.

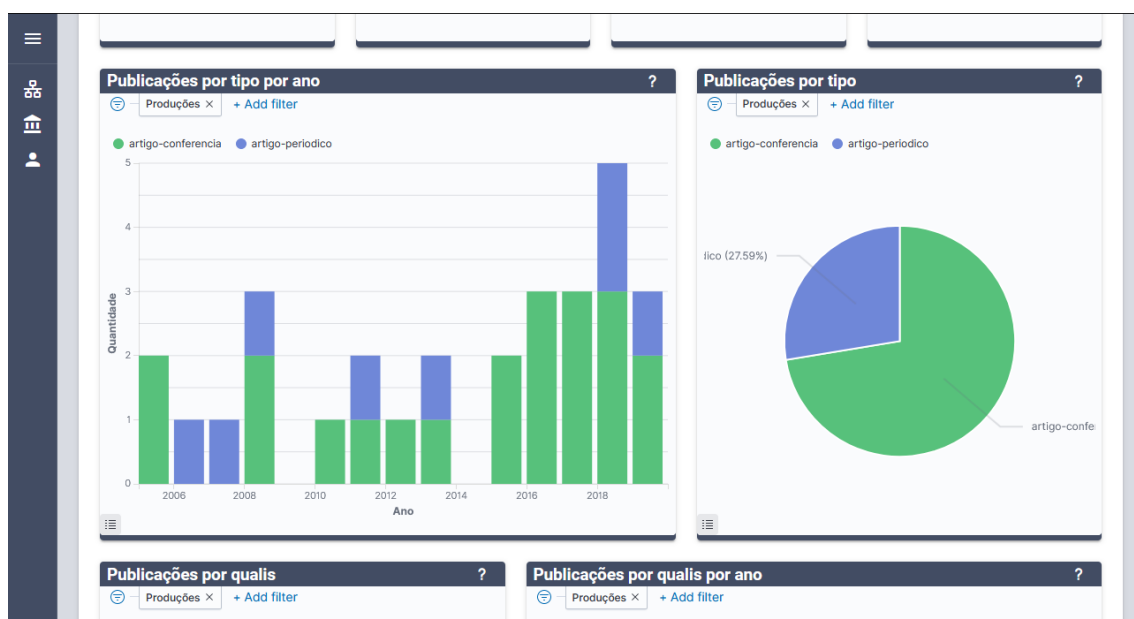


**Figura 16 – Início da funcionalidade - seleção dos pesquisadores que irão compor a rede**

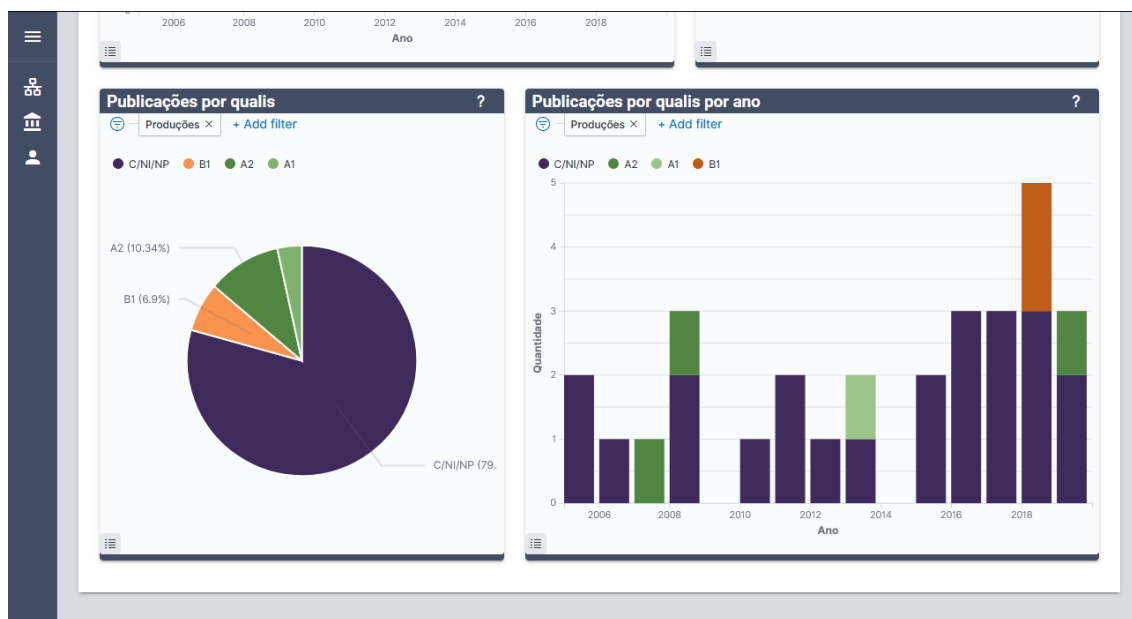




**Figura 17 – Métricas sumarizadas da rede de colaboração**



**Figura 18 – Gráficos da quantidade de publicações por tipo no tempo e quantidade por tipo no geral**



**Figura 19 – Gráfico da publicação por qualis, geral e por período**

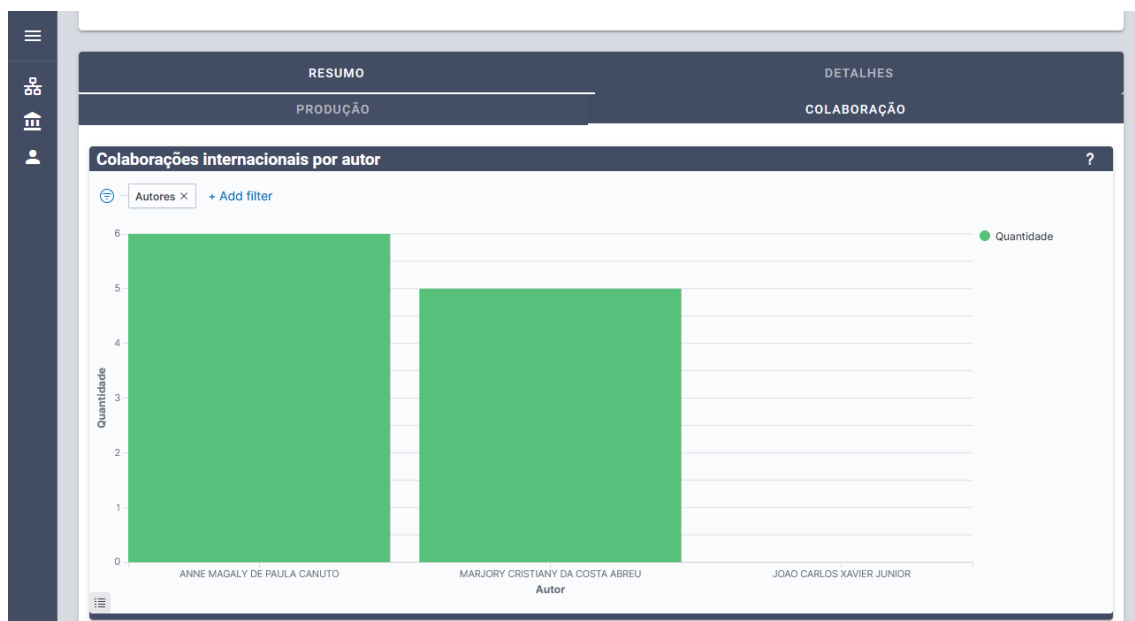
#### **4.3.3. Análise da rede de colaboração (resumo - colaboração) - informações exibidas**

- número de colaborações internacionais para cada pesquisador;
- média de pesquisadores em colaborações internacionais para cada pesquisador.

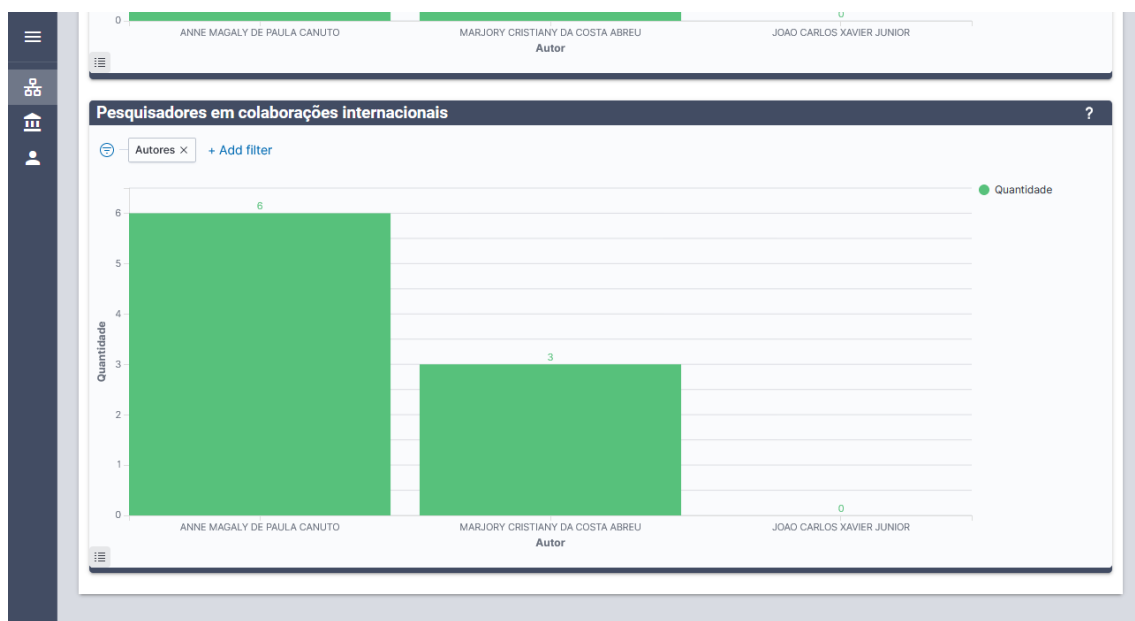
#### **4.3.4. Análise da rede de colaboração (resumo - colaboração) - capturas de tela**

A Figura 20 mostra um gráfico com o número de colaborações internacionais de cada pesquisador que constitui a rede. No eixo horizontal temos os pesquisadores e no vertical temos a quantidade de publicações.

A Figura 21 mostra o número de pesquisadores em colaborações internacionais com a participação de cada membro da rede. No eixo horizontal temos os pesquisadores que constituem a rede e no vertical a quantidade de colaboradores internacionais que atuaram com os pesquisadores.



**Figura 20 – Gráfico das colaborações internacionais dos pesquisadores da rede**



**Figura 21 – Gráfico dos pesquisadores em publicações internacionais da rede**

#### 4.3.5. Análise da rede de colaboração (detalhes) - informações exibidas

As informações exibidas nessa parte são repetidas para cada sub-rede. Por exemplo, se uma rede é composta pelos pesquisadores *A*, *B* e *C*, esta parte da funcionalidade irá exibir as informações para as sub-redes *A-B*; *A-C*; *B-C*; e *A-B-C*.

1. número de publicações por qualis;
2. número total de publicações;
3. citescore das publicações.

#### 4.3.6. Análise da rede de colaboração (detalhes) - capturas de tela

A Figura 22 mostra dois gráficos. O primeiro exibe a porcentagem de cada qualis em relação ao total de publicações da rede. O qualis que cada fatia representa pode ser consultado na legenda. O segundo diz respeito à relevância das produções, trazendo a proporção entre produções de grau alto, médio e baixo de relevância dentre as publicações da rede.

A Figura 23 mostra a quantidade de produções por pesquisador. No eixo horizontal temos os pesquisadores e no eixo vertical o número de produções.

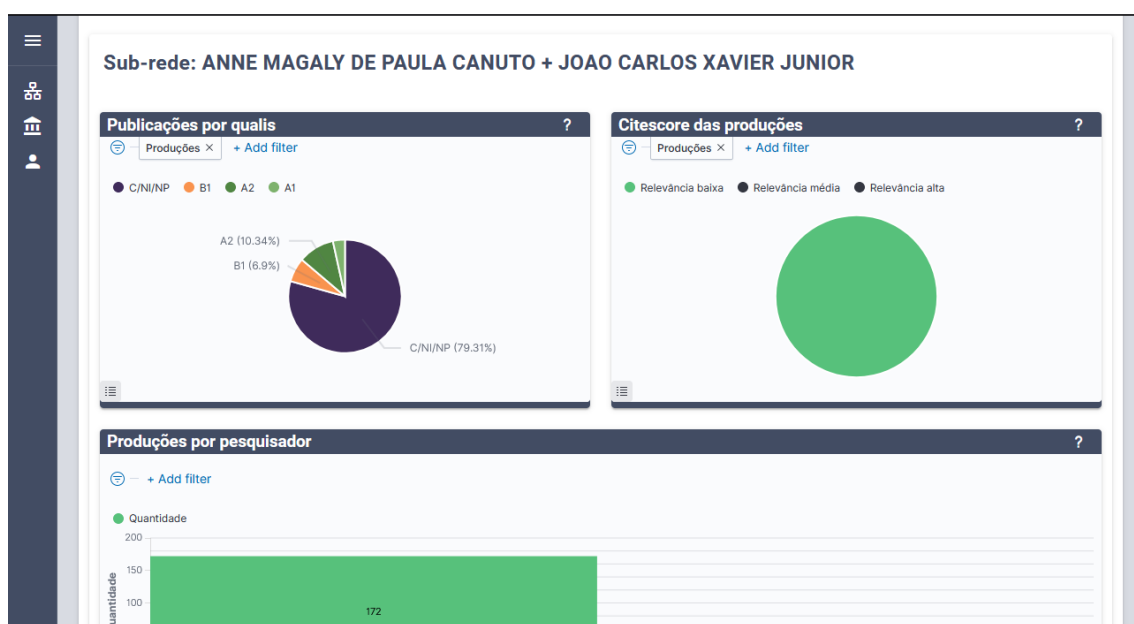


Figura 22 – gráficos das publicações por qualis da sub-rede analisada

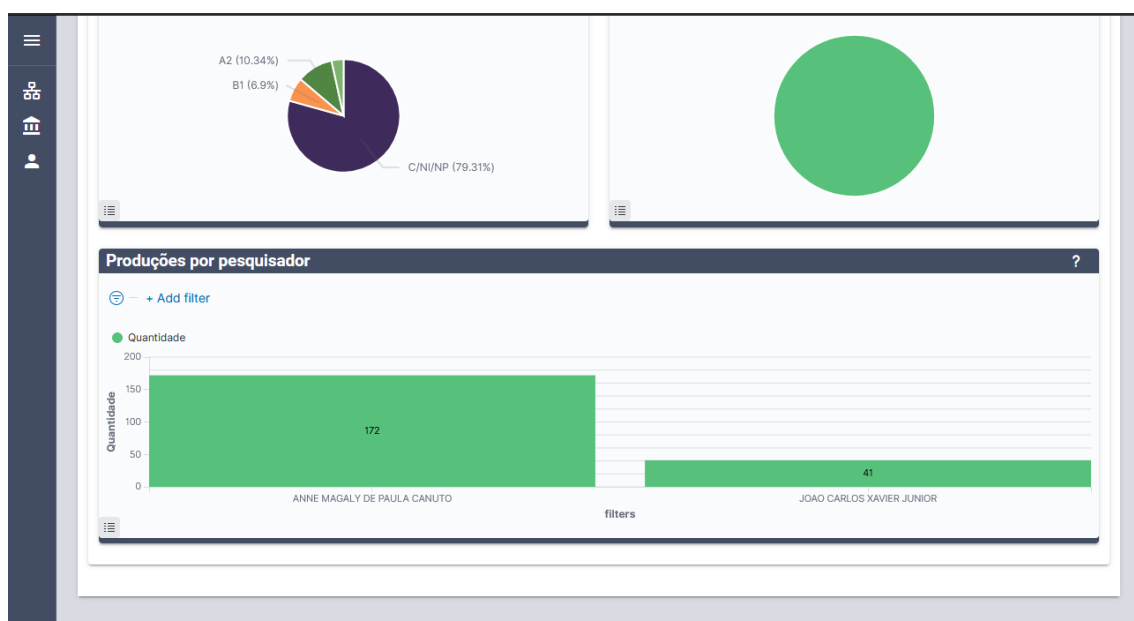


Figura 23 – gráficos das publicações por qualis da sub-rede analisada

## 5. Considerações Finais

O presente relatório apresentou o protótipo de sistema desenvolvido para demonstrar o funcionamento das seguintes funcionalidades: análise de pesquisadores com base em suas pesquisas; análise da produção científica de programas de pós-graduação; e análise da produção científica de redes de colaboração. Além da descrição geral de cada funcionalidade, foram citados os dados usados para construção, informações que são exibidas e capturas de tela para cada funcionalidade, todos seguindo o fluxo visual presente no sistema desenvolvido.

De maneira geral, é possível afirmar que tratou-se de uma pesquisa desafiadora, dado que a análise de produção científica sempre está rodeada por questões relacionadas à inviabilidade de comparar possíveis entes de contextos diferentes, questões relacionadas ao tempo em que foi feita a pesquisa e etc. Contudo, a utilização de técnicas não-supervisionadas de Aprendizado de Máquina possibilitou a segregação adequada dos dados e, como consequência, a observação dos entes individuais frente aos grupos gerados, seja para a funcionalidade de análise de pesquisadores ou mesmo para a análise de programas de pós-graduação.

Em relação à última funcionalidade apresentada neste relatório, a análise de redes de colaboração, é necessário mencionar a dificuldade de tal análise devido à complexidade computacional presente no cruzamento de dados de produções de pesquisadores em um universo de busca tão grande. Essa dificuldade foi mitigada com a utilização de dados intermediários construídos sob demanda e mantidos a cada consulta.

Foram utilizadas as seguintes tecnologias em toda a pesquisa:

- **análise e exploração de dados:** Weka-API, Weka-GUI, Scikit-learn, Pandas, Numpy;
- **back-end, consultas e gráficos:** Elastic Search, Kibana;
- **front-end:** Vue-js.

É importante também citar que, mesmo no sistema protótipo, levamos em consideração a dificuldade de entendimento do funcionamento de cada funcionalidade e também das possibilidades de personalização dos gráficos em cada tela. Por esse motivo foram adicionados campos de ajuda, com explicações detalhadas sobre cada funcionalidade e também sobre os gráficos.

Por fim, todas as bases, códigos, este relatório e também o vídeo de apresentação da plataforma em utilização estão presentes no repositório [https://github.com/ml-imd/Sclx\\_research](https://github.com/ml-imd/Sclx_research). Nele os diversos artefatos estão separados por pastas.