Atividade: Relatório de evolução de desenvolvimento de ferramentas de monitoramento, análise e visualização de dados do CGEE

Exploração de dados e visualização de informação

Atividade: Relatório de evolução de desenvolvimento de ferramentas de monitoramento, análise e visualização de dados do CGEE

Exploração de dados e visualização de informação

Brasília, DF
dezembro, 2021

Centro de Gestão e Estudos Estratégicos (CGEE)

Organização social supervisionada pelo Ministério da Ciência, Tecnologia e Inovações (MCTI).

Relatório de evolução de desenvolvimento de ferramentas de monitoramento, análise e visualização de dados do CGEE. Exploração de dados e visualização de informação. Brasília: Centro de Gestão e Estudos Estratégicos, ano 2020.

179p.: 130 il.

1. Ciência de dados. 2. Análise de redes complexas. 3. Ciência e Tecnologia I. CGEE. II. Título.

Centro de Gestão e Estudos Estratégicos (CGEE), SCS Qd 9, Torre C, 4º andar, Ed. Parque Cidade Corporate, CEP: 70308-200 - Brasília, DF, Telefone: (61) 3424 9600, http://www.cgee.org.br

Referência bibliográfica:
Centro de Gestão e Estudos Estratégicos- CGEE. Relatório de evolução de desenvolvimento de ferramentas de monitoramento, análise e visualização de dados do CGEE. Brasília, DF: 2020. 134p.

Este relatório é parte integrante das atividades desenvolvidas no âmbito do 2º Contrato de Gestão CGEE – 11º Termo Aditivo. Programa: (Exploração de dados e visualização de informação). Projeto – (8.10.56.01.51.01).

Atividade: Relatório de evolução de desenvolvimento de ferramentas de monitoramento, análise e visualização de dados do CGEE

Exploração de dados e visualização de informação

Supervisão
Marcio de Miranda Santos

Coordenador
Jackson Max Furtunato Maia

Equipe técnica do CGEE
Alberto Akira Okata
Amanda Queiroz Sena
César Augusto Costa
Eduardo Amadeu Dutra Moresi
Evandro Augusto Soares
Genilda Carlos da Mota
Ícaro Lorran Lopes Costa
Israel Garcia de Oliveira
Kleber de Barros Alcanfôr
Marcus Vinícius Tavares da Cunha Mello Filho
Rogério da Silva Castro

Consultor
Jörg Neves Bliesener

Sumário

1. Introdução

O reconhecimento e análise de informações existentes nas grandes massas de dados atualmente acessíveis permitem multiplicar a capacidade de atuação do CGEE, desde que técnicas adequadas de extração, tratamento e carga de dados sejam empregadas para reconhecer padrões que lhes sejam subjacentes. Nesse sentido, o projeto "Exploração de Dados e Visualização de Informações" visa fortalecer as competências do Centro, desenvolvendo e validando fundamentos, metodologias e ferramentas de análise de dados de CTI disponíveis, ampliando seu portfólio de serviços e auxiliando o embasamento metodológico das suas demais atividades e ações.

2. Análise de redes de documentos e de currículos Lattes

Nos últimos anos o CGEE consolidou competências na análise de redes complexas aplicada a bases de dados relevantes para Ciência, Tecnologia e Inovação (CTI). Ao longo das implementações em software dos processos de análise, foi percebida a conveniência de uma separação entre a fase de análises exploratórias de dados, voltada para a mineração de dados e reconhecimento de padrões, e a fase de comunicação dos dados minerados e padrões reconhecidos. Para análises exploratórias, as interfaces de usuário devem ser elaboradas para facilitar o trabalho do **analista**, em princípio um especialista no domínio de conhecimento relacionado ao problema, com visualizações que simplifiquem o máximo possível a complexidade matemática, estatística ou computacional do tratamento de dados. Na segunda fase é mais importante que o esforço de desenvolvimento seja concentrado na **plateia** para a qual os resultados deverão ser comunicados, ou seja, o cliente do CGEE, de modo que os requisitos devem se concentrar mais na apresentação de resultados do que na sua exploração.

Essa separação definiu ferramentas com as duas funções correspondentes. Com o InsightNet (iN), o usuário realiza o trabalho de análise e preparação dos arquivos de rede a serem exportados para posterior exploração visual em outra ferramenta, o InsightNet Browser (iNB). Posteriormente, algoritmos de *back end* desenvolvidos para o iN foram incorporados à ferramenta de mineração textual InsightData (iD), de modo que seu acervo também pode ser analisado em redes de similaridade semântica de documentos no iN e seus resultados exportados para o iNB. Com esta conexão entre suas três principais ferramentas de análise de textos e metadados associados, o Centro tem a possibilidade de realizar análises de redes e mineração de dados com documentos do acervo próprio (integrado ao iD), de currículos Lattes e de metadados de algumas das principais bases de dados de produção acadêmica e de patentes disponíveis, além de textos e metadados das suas consultas e outras bases de dados semiestruturados. Como atividades centradas no aprimoramento das ferramentas iN, iNB e metodologias associadas, destacam-se:

a) Lançamento da versão 3.2.10 da ferramenta iN com: atualização para uso do novo *web service* do CGEE (baseado no banco de dados PostgreSQL), implementação de um novo tratamento de duplicatas em registros das bases de dados bibliográficos, além de melhorias nos formatos de saída para o iNB que permitirão análises simultâneas de metadados das

redes geradas. Essa melhoria permitirá análises muito mais abrangentes, pois permitirá que padrões identificados em um determinado conjunto de dados dispostos como uma rede de relações, possão ser comparados em outras redes que contenham dados relacionados aos primeiros, incluindo seus metadados. O manual atualizado da ferramenta pode ser consultado no Apêndice deste relatório.

b) Lançamento da versão 1.7.0 do iNB, incluindo correções de bugs e o tratamento dos novos formatos de saída do iN para permitir as análises simultâneas das redes geradas, mencionadas no item anterior. A funcionalidade básica que permite essa nova abordagem é tratamento de metadados com campos do tipo lista para geração de gráficos de contagem de nós, agrupados pelos elementos da lista. Oberve-se que, como as modificações afetam apenas os processos internos de leitura e transformações dos dados, não houve a necessidade de atualizar o manual.

c) Foi feito também o lançamento de uma primeira versão web do iNB que, assim, inaugura um novo modelo de disponibilização de dados *online* para usuários externos e viabiliza a utilização de recursos não disponíveis para a versão desktop usual, como acesso a banco de dados dedicado, acesso a múltiplos usuários à mesma base de dados e possíveis ganhos de desempenho associados ao maior poder de processamento dos servidores do Centro. Já aproveitando o uso de banco de dados do servidor, a parte de carga do código foi refatorada para perimtir o tratamento de redes consideravelmente maiores (em número de nós e/ou arestas) do que a versão desktop, que é vinculada às limitações do browser.

d) Como uma evolução natural dos métodos desenvolvidos para as ferramentas iN, iNB e iD, ao longo de 2021, foram discutidos entre os membros da equipe EDVI os primeiros resultados de implementação em Python e testes dos protótipos de *back end* e *front end* de um conjunto de soluções baseadas nos algoritmos já empregados na iN (originalmente escrito em Java), para cálculo e visualização de redes (inclusive 3D) com volumes de dados pelo menos 10 vezes maiores que a versão original. As implementações da versão em produção dessas redes de *big data* (juntamente com outras soluções e novos indicadores bibliométricos desenvolvidos com métodos desenvolvidos fora do contexto do Projeto EDVI) foram aplicadas e tiveram sua evolução validadas no contexto do Projeto de Atividade do Observatório de Ciência, Tecnologia e Inovação (OCTI). Os principais resultados dessas aplicações podem ser encontrados na página https://octi.cgee.org.br/. Uma versão para uso

geral por parte de outras equipes do Centro está sendo desenvolvida pela equipe da TI.

e) Vale notar foi dada continuidade à profícua sinergia entre a equipe do projeto e a equipe de tecnonologia da informação (TI) do Centro, aprofundada em 2020. Neste ano, houve um significativo aumento da interação entre as equipes com foco na utilização otimizada dos ambientes de implantação, versionamento e distribuição, facilitando o acesso dos usuários finais às versões mais recentes dos softwares em intervalos de tempo bem mais curtos entre o desenvolvimento e produção. Além disso, a experiência de uso dos ambientes disponibilizados para o iN e o iNB motivou a equipe do projeto a iniciar a migração todos os protótipos para este modelo de fluxo de trabalho.

3. Novos algoritmos e protótipos

Ao longo dos últimos 7 anos, o CGEE tem se aprofundado no desenvolvimento de algoritmos e ferramentas de análise de dados relacionais, particularmente com base em técnicas, indicadores e formalismos de teoria de redes complexas junto com conceitos e métodos do processamento de linguagem natural e da recuperação da informação.

Começando em 2013, as aplicações de análises relacionais, ou análises de redes, foram implementadas em códigos escritos em linguagem Java, para aproveitar a universalidade de plataformas que podem executar seus códigos e a disponibilidade de ambientes de visualização nessa linguagem. Para acelerar o processo de desenvolvimento, os esforços de codificação foram centrados na camada de acesso e tratamento dos dados (*back end*) de maior interesse. A ferramenta iN, cujo *back end* foi quase todo originalmente desenvolvido no Centro para ser executado aproveitando os recursos de *front end* da ferramenta Gephi, e que tem quase todos os seus códigos dedicados à carga e transformação de dados bibliométricos e conectados a um espelho da base de dados Lattes é o exemplo mais bem-sucedido dessa abordagem de desenvolvimento.

A partir de 2016, novas abordagens foram implementadas para suprir novas demandas que surgiram. Além de novas necessidades de tratamento de bases de dados textuais diferentes das incluindas no iN e iD, as visualizações fornecidas pelo aplicativo Gephi já não eram suficientes, do ponto de vista de acabamento, para as entregas demandadas pelas várias equipes de projetos diferentes. Além disso, o modelo de entrega de produtos de projetos com visualizações com imagens estáticas em relatórios ou apresentações também foi se tornando um importante limitador das possibilidades de apresentação de resultados dos estudos contratados, que gradativamente requeriam visualizações interativas confluentes às práticas de navegação por browser comuns a qualquer usuário moderno da internet.

Para o modelo de entrega de produtos com *front end* de análise visual interativa de dados foi desenvolvida internamente a já comentada ferramenta iNB. Esse aplicativo é compatível com o conceito mais recente na área de visualização da informação chamada *visual analytics environment*, ambientes nos quais usuários experientes nos domínios da informação apresentada podem fazer consultas e raciocínios aplicando filtros visuais (cores, formas ou grafos, por exemplo, representando os dados). A ferramenta, baseada em uma

proposta inicial disponibilizada na página do Gephi, foi totalmente refatorada e está implementada em JavaScript.

Juntamente com outras iniciativas na direção de estudos cada vez mais baseados em evidências, essas ferramentas contribuíram para expandir a capacidade analítica do CGEE. Como consequência bem-vinda dessa cultura institucional crescentemente baseada em dados, novos projetos do Centro trouxeram novas demandas de processamento de dados estruturados e não estruturados, numéricos e categóricos e de fontes não usuais como as tradicionais plataformas Lattes ou Sucupira (provida pela Capes). Essas demandas já não eram compatíveis com a capacidade de desenvolvimento instalada para as ferramentas do pacote *insight*. Para supri-las, foram realizadas novas contratações em 2019 e 2020 e, com a incorporação das competências da nova equipe, foram testadas modificações na filosofia de desenvolvimento no sentido de prototipagem mais rápida de ideias, tentando reter características de usabilidade dos protótipos desenvolvidos para usuários experientes nas temáticas estudadas, mas sem perfil técnico em matemática, estatística ou computação.

Após testes de protótipos desenvolvidos na linguagem R e Python para o *back end* e JavaScript para o *front end,* foi definido por consenso da equipe um novo fluxo de desenvolvimento. Segundo esse fluxo, dado um problema de dados e escolhidos ou desenvolvidos algoritmos adequados para sua solução, seus códigos são inicialmente validados em *notebooks* do Jupyter, um ambiente de execução de códigos de programação que simplifica bastante as frequentes alterações comuns em ciência de dados. Uma vez validados no contexto da equipe EDVI, os programas são integrados a interfaces WEB baseadas em JavaScript com o devido planejamento de protótipos de visualizações. As aplicações resultantes desse processo são então disponibilizadas paras os demais usuários do CGEE, para nova etapa de validação, seja das soluções implementadas no *back end*, seja nas interfaces de usuário dos códigos escritos para o *front end*. Os protótipos mais bem aceitos pelos usuários do Centro deverão então ser entregues à TI para novos avanços nas suas maturidades tecnológicas na direção de produtos acabados.

Em uma outra vertente de desenvolvimento viabilizada pelos novos membros da equipe, nos últimos dois anos, técnicas de aprendizado de máquina, particularmente redes neurais artificiais, também têm sido testadas em algumas aplicações em diferentes projetos do Centro.

Em 2020, várias aplicações foram desenvolvidas ou aprimoradas de acordo com essas novas linhas de desenvolvimento e foram descritas no relatório pertinente. Esse desenvolvimento foi focado na ambrangência de problemas a serem resolvidos, de acordo com demandas específicas de projetos. Em 2021, a equipe se concentrou mais em aprimorar as ferramentas anteriormente lançadas, sem perder de vista algumas novas aplicações. No que segue, serão resumidos os protótipos que, mesmo em diferentes estágios de maturidade, chegaram pelo menos a uma etapa do processo que permite primeiros testes por parte dos usuários finais do CGEE. Também são relatados avanços em aplicações de inteligência artificial para a classificação de documentos de interesse, embora nesses casos buscou-se não uma ferramenta para usuários em geral, mas um protótipo de finalidade específica para as necessidades de um projeto em particular. Em 2021, foram desenvolvidos ou aprimorados os seguintes protótipos:

a) Extrator de dados de patentes registradas em currículos Lattes – O CGEE tem acesso a fontes de dados públicos de escritórios de patentes. Cada escritório tem seu próprio formato de dados e o tratamento de cada fonte demanda um processo diferente de extração. Para estudos do potencial de inovação de grupos de pesquisa ou de pesquisadores, tem sido útil considerar os dados registrados na plataforma Lattes. Aproveitando parte de um desenvolvimeto de 2020, o editor de arquivos XML, foi desenvolvida uma ferramenta de extração de dados de *tags* de lotes de arquivos XML (como um conjunto de currículos obtidos por busca por CPF, por exemplo) e exportação dos dados extraídos em formatos tabulares (CSV, Excel), para análise posterior, inclusive no próprio iN. A principal aplicação do protótipo é a geração de planilhas com dados de patentes ou outras produções tecnológicas, mas outras informações contantes no Lattes podem ser extraídas pela ferramenta, como lista de orientandos, lista de projetos de pesquisa e vínculos empregatícios, para citar apenas alguns usos possíveis. A Figura 1 mostra a tela inicial da nova funcionalidade de extração de dados de arquivos XML:

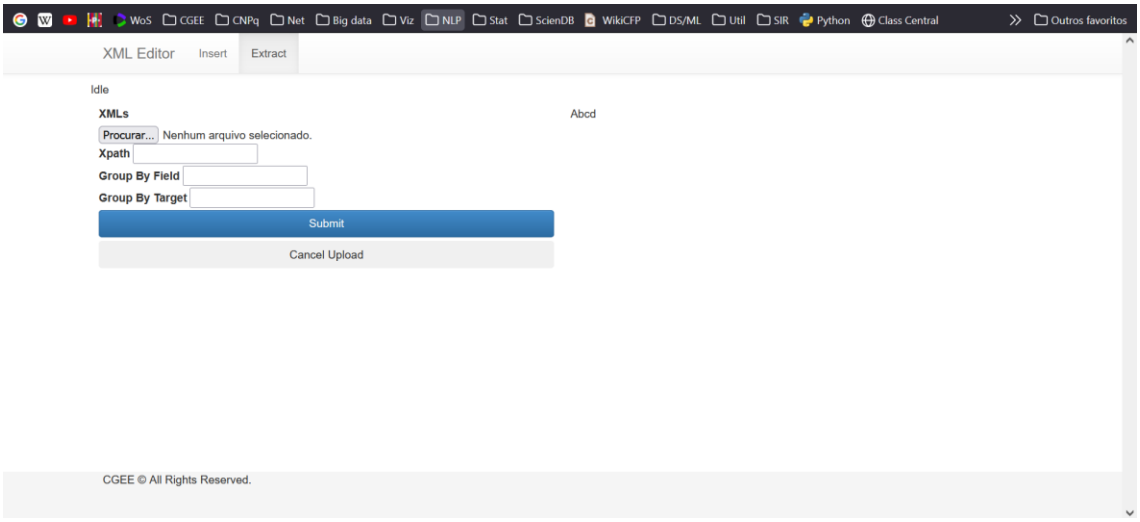


Figura 1: Tela de entrada do extrator de XML

b) Testes de algoritmos para a transcrição de áudios – Vários projetos do CGEE envolvem reuniões de especialistas ou a participação de funcionários em reuniões. As transcrições das conversas, quando gravadas, são feitas por empresas especializadas. Com o objetivo que contar com uma solução local de para a transcrição automática de gravações de reuniões, particularmente reuniões online que facilitem a produção de atas, alguns algoritmos de reconhecimento automáticos de fala baseados em redes neurais artificiais foram testados. Infelizmente, as bases de arquivos de áudio gratuitas para o treinamento dos algoritmos não têm bons exemplos de português brasileiro e as transcrições resultantes não tinham boa qualidade. Esta inciativa foi, portanto, suspensa até encontrarmos bases de áudio de português brasileiro de melhor qualidade.

c) Ferramenta para segmentação automática de textos – A segmentação automática de textos visa dividir um texto em partes semanticamente significativas. Esses segmentos podem ser separados em tópicos, sentenças ou mesmo em palavras. Como, em reuniões, diferentes tópicos podem ser abordados por falantes diferentes em instantes diferentes, a organização de trechos que se referem a um mesmo tópico pode facilitar um resumo de reunião ou uma ata se sentenças similares forem classificadas e agrupadas em seus tópicos particulares. Para essa finalidade foi desenvolvida uma extensão do sumarizador de textos que utiiza a técnica de similaridade semântica no nível de parágrafo e permite uma anotação automática de trechos provavelmente similares que constituem tópicos. Um exemplo de uso da ferramenta é mostrado na Figura 2 abaixo:

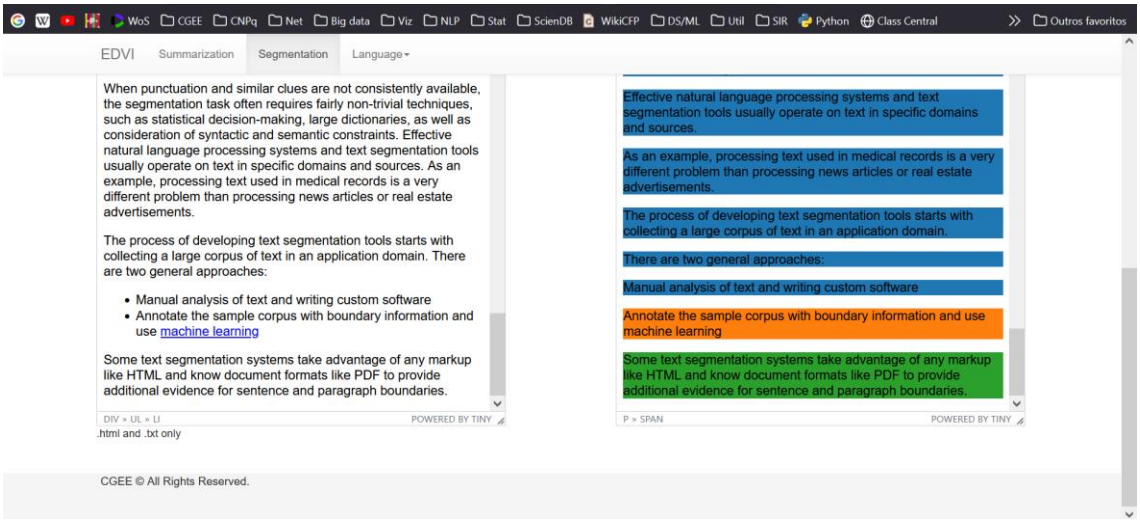


Figura 2: Exemplo da ferramenta de segmentação de textos. Cada cor corresponde a tópicos em princípio diferentes.

d) Redes de similaridade entre palavras – As redes de similaridade semântica de textos são compostas por nós que representam os textos conectados por arestas que representam o grau de similaridade entre um par de textos. Quando existem muitas arestas conectando vários pares de textos, existe uma tendência à formação de *clusters* semânticos que são extremamente úteis nas análises realizadas, pois normalmente cada *cluster* define um domínio temático, ou tópico, de tal forma que essa técnica pode ser considerada um tipo de modelagem de tópicos. Mas, em comparação a abordagens tradicionais de modelagem de tópicos, é possível extrair mais informação com a abordagem de redes de similaridade.

Para caracterizar o conteúdo temático do *cluster*, as ferramentas de análise desenvolvidas no CGEE coletam as palavras-chave por *cluster* e um escore de frequência dessas palavras-chave quase sempre é suficiente para sua classificação temática. Entretanto, em conjuntos de dados que não contêm palavras-chave, ou conjuntos nos quais estas não fornecem resolução semântica suficiente, a classificação temática é comprometida e o analista tem que ler os textos para realizar uma classificação manual. A ferramenta desenvolvida em 2020 coleta os termos mais relevantes na constituição dos pesos das arestas que conectam pares de textos. Quando analisados os *clusters* de documentos, os escores de relevância de cada termo computados para todas as arestas do *cluster* ajudam a determinar o seu domínio temático de uma forma significativamente mais efetiva, mesmo nos casos que os textos têm palavras-chave entre seus metadados.

Uma evolução natural do aplicativo para mineração desses termos conectores é calcular a rede inversa de similaridade entre os prórpios termos, calculadas a partir dos textos nos quais eles são mais relevantes. Essa funcionalidade foi desenvolvida tendo como meta final

a geração de taxonomias que representem o *corpus*. Os primeiros resultados mostraram que ainda é necessário algum trabalho manual de limpeza de termos irrelevantes e será buscada uma maior automatização nessa seleção. Na etapa atual de desenvolvimento, redes com os nós sendo termos do *corpus* são geradas e exportadas para visualização e análise e edição manual no iN. Um exemplo da visualização fornecida no iN a partir de dados da ferramenta é exibido na Fig. 3.

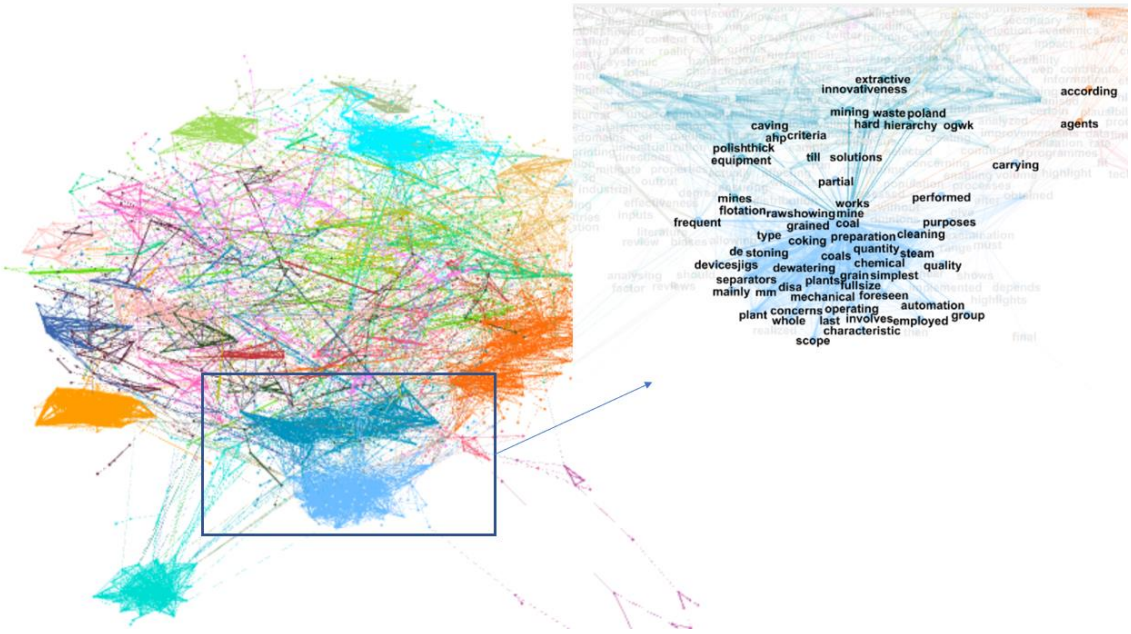


Figura 3: Exemplo de rede de similaridades entre palavras de um corpus construída a partir de dados preparados pela ferramenta descrita no item d) acima.

e) Ferramenta para a edição de conjuntos de dados semiestruturados em formato CSV – Em muitos casos, a análise de conjuntos de dados bibliográficos demanda que subconjuntos de dados sejam analisados em separado. Embora a edição desses subconjuntos posso sar feita manualmente, para milhares de registros, essa tarefa é inviável. Uma ferramenta para edição de arquivos CSV foi desenvolvida para essa finalidade. Essa ferramentfoi inicialmente projetada para arquivos CSV do *Web of Science,* mas pode ser facilmente preparada para outras fontes de dados nesse formato (como Derwent, Scopus). A sua aplicação ocorre após a seleção dos subconjuntos de registros dentre os do conjunto completo pelo uso dos filtros existentes seja no iN, seja no iNB. Esses programas também permitem a exportação de listas de identificadores de cada um dos registros selecionados. Com essa lista e o conjunto de dados original, o protótipo reconfigura o arquivo original e gera um novo CSV apenas com os registros da lista selecionada. A Figura 4 mostra a tela de entrada da ferramenta.

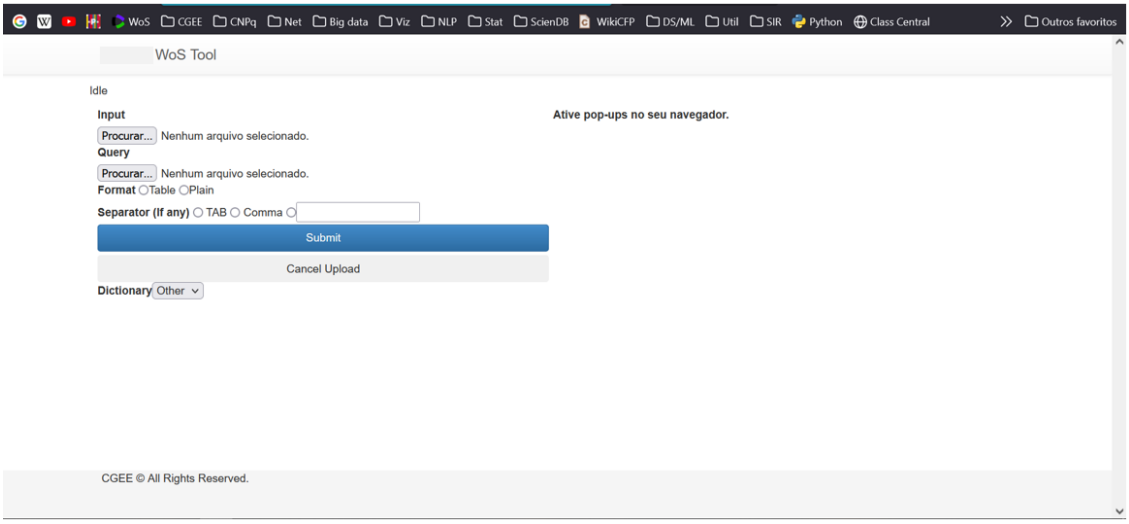


Figura 4: Ferramenta para manipulação de conjuntos de dados CSV (arquivos de Input, na figura) a partir de um arquivo de filtro (Query, na figura).

f) Evolução na ferramenta de visualização de dados categóricos (+Tab) – O +Tab foi concebido em 2020 apenas para a visualização de mapas de calor de coocorrências e probabilidades condicionais entre dados categóricos de planilhas resultantes de consultas públicas realizadas na plataforma insightSurvey. Em 2021 esse protótipo foi refatorado para melhorar seu desempenho na leitura de dados e incorporar novas funcionalidades.

Uma nova funcionalidade é a possibilidade de, em uma tabela que tenha colunas de dados numéricos, transformá-los em categóricos, agrupando-os por faixas de valores, sendo que o número de faixas é definido pelo usuário. Dessa forma, pode-se analisar as coocorrências entre o número de objetos por faixa e as demais variáveis, como, por exemplo, o número de artigos em uma determinada faixa de número de citações por país.

Além disso, na versão de 2021, a ferramenta foi repensada para, além de permitir a visualização das coorrencias de valores de uma dada variável categórica, também servir de base para o tratamento de redes *multilayer* de coorrências. Nesse tipo de rede, cada camada corresponde a uma rede tradicional, mas há conexões entre os nós das diferentes camadas. Como os mapas de calor do +Tab já eram visualizações naturais de matrizes de supra-adjacências (*supra-adjecency matrix*, ou SAM), as representações matriciais de redes *multilayer,* a ferramenta foi reestruturada para exportar as diferentes camadas como redes específicas. O conjunto de arquivos de redes podem, ser tratados no iN e, posteriormente, consolidados como janelas diferentes no iNB, onde as arestas entre as camadas podem ser examinadas pelo uso da funcionalidade de busca deste programa. Como um primeiro exemplo de aplicação dessa técnica, foi desenvolvida uma nova proposta metodológica de mapeamento temático de patentes a partir das redes de

coocorrência entre códigos de classificação de patentes (como o IPC ou o CPC). Essa proposta ainda está em testes, mas o conceito pode ser expandido para a maioria dos tipos de análises que são feitas no CGEE. Um exemplo apenas para ilustração de mapas de calor (que representam uma SAM) contendo duas camadas é mostrado na Figura 5.

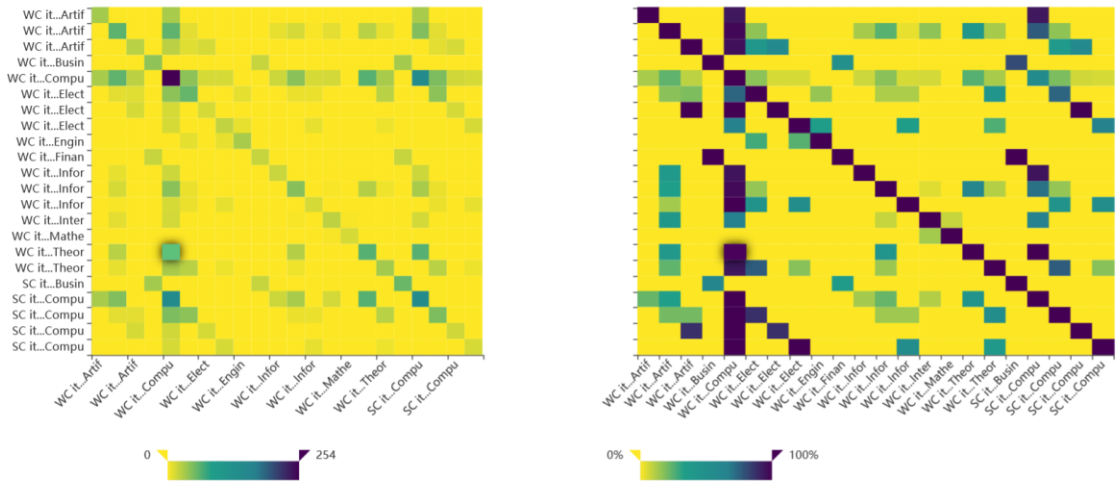


Figura 5: Mapas de calor mostrando as ocorrências entre os valores de duas varáveis. A figura da esquerda representa as coocorrências e a figura direita representa o conjunto de probabilidades condicionais entre as coocorrências entre as variáveis.

g) Ferramenta para análise de dados numéricos (+Tab_n) – A ferramenta descrita no item anterior busca identificar padrões em coocorrências, típicos de dados categóricos. Em análise multivariada de dados, um conjunto de *m* objetos a serem analisados são representados por sequências de *n* números, onde cada número corresponde ao valor de uma variável aleatória. Esse conjunto pode ser representado por uma tabela de números.

Em muitos casos, padrões nesses dados numéricos podem ser obtidos pelo cálculo de relações entre os objetos (linhas da tabela) calculadas a partir das n-uplas de números (conjunto de valores de cada coluna para aquela linha) que os caracterizam. Se essas relações forem agrupadas sob a forma de redes de relações, os métodos de análises de redes tradicionalmente empregados no CGEE também podem ser aplicados.

O objetivo do +Tab_n é, portanto, estender técnicas desenvolvidas para o +Tab para casos onde os dados são numéricos sendo que, no lugar de redes de coocorrências, o protótipo gera redes de relações numéricas entre as linhas da tabela, par a par. Uma primeira versão da ferramenta foi desenvolvida e testada com cálculos de 4 tipos de correlações entre as linhas da tabela: correlações de Pearson, Spierman, phi_k e relações de similaridade de cosseno. Tomadas em conjunto, essas relações têm a intenção de facilitar e tornar visuais (com suas respectivas redes) parte do processo de análise exploratória de dados

numéricos, em complemento a outros passos da análise exploratória, como estatísticas descritivas e cálculos de distribuições e de *outliers*.

O processo de preparação começa com a carga da tabela (em arquivo CSV ou planilha do Excel). O programa permite uma pré-visualização dos dados como um mapa de calor para ajuda o usuário a conceber estratégias de exploração, de acordo com a figura abaixo:

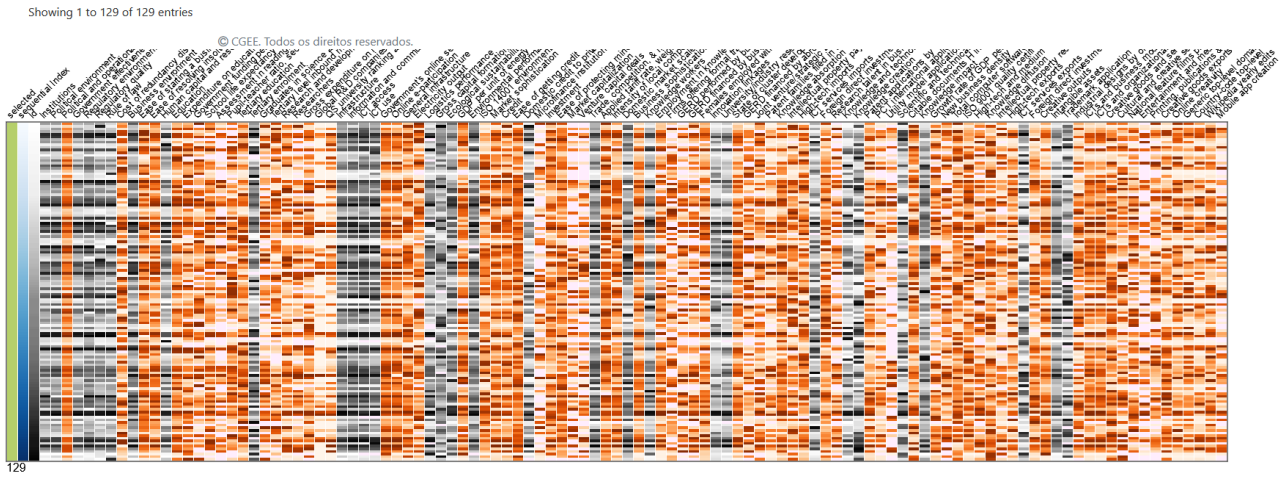


Figura 6: Mapa de calor de uma tabela onde as linhas correspondem a países e as colunas são os 80 indicadores do índice global de inovação. A intensidade de cor corresponde a valores entre 0 e 100.

Na segunda etapa, o usuário escolhe qual dos 4 índices de correlação deseja empregar para estabeler os pesos das arestas que conectam cada par de objetos (as linhas e colunas), representados também como intensidade de cor nos cruzamentos entre linhas e colunas. Um novo mapa de calor (ou matriz de adjacência da rede) é gerado a partir das correlações:

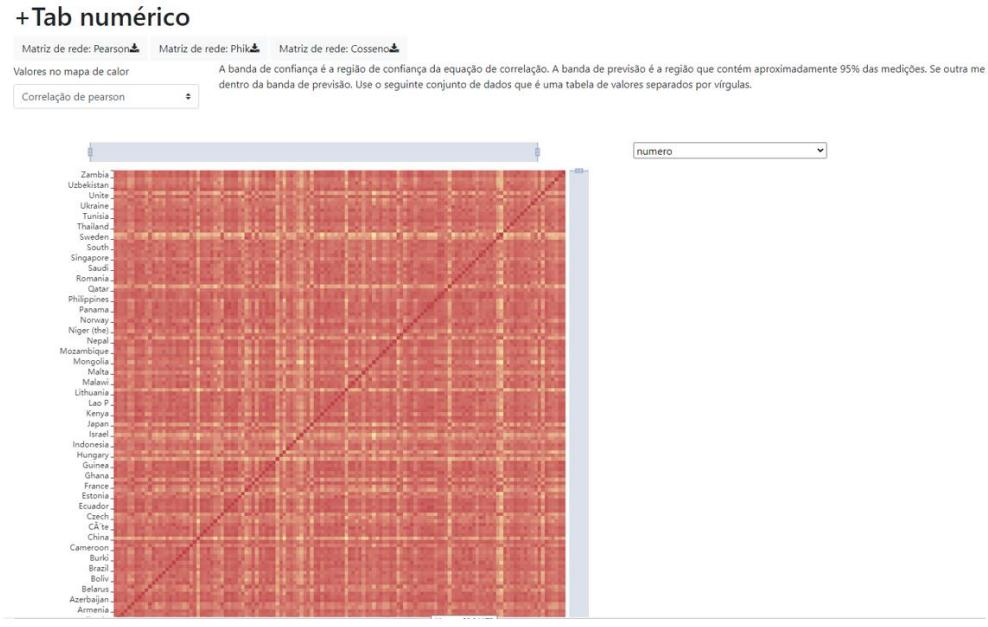


Figura 7: Mapa de calor das correlações entre os conjuntos de valores dos 80 indicadores de cada país, usando o mesmo exemplo da figura anterior. Note que na diagonal as correlações são máximas, 1, porque representam as correlações entre cada país e ele mesmo.

O mapa de calor das correlações pode ainda ser examinado com mais detalhes clicando em cada um dos cruzamentos entre linhas e colunas para uma explicitação dos dados que compõem a correlação (ver Figura 8):

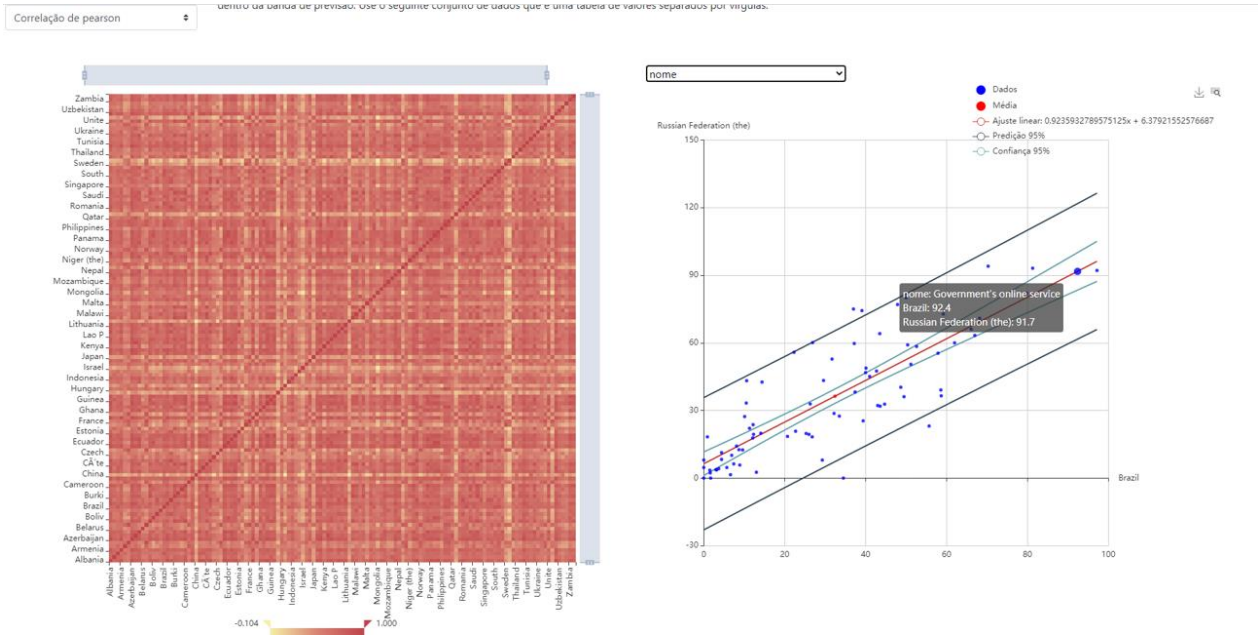


Figura 8: explicitação de um dos pontos do mapa de calor da figura anterior ressaltando a correlação entre o Brasil e a Rússia. Os pontos do gráfico correspondem aos escores dos dois páises para cada indicador do índice global de inovação, além dos parâmetros estatísticos do coeficiente de correlação empregado.

Por fim, o grafo final, com as linhas representando os nós e as arestas representando os coeficientes de correlação do mapa de calor da figura anterior pode ser exportado para mais análises (Figura 9):

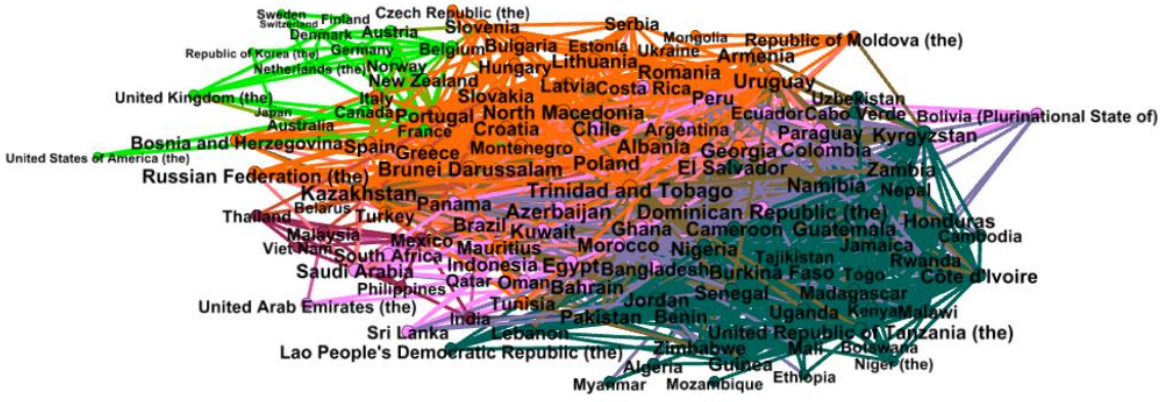


Figura 9: Rede de correlações entre escores de indicadores do índice global de inovação dos países.

h) Prova de conceito de um "empacotador" de versões desktop das aplicações desenvolvidas – O desenvolvimento de protótipos deste projeto segue a diretriz geral de que estest devem ser aplicações web executadas em um servidor e operadas pelo usuário via browser. Entretando, antecipando a possibilidade de haver demandas de execução dos programas nos computadores dos próprios usuários, foi desenvolvido um conjunto de programas para “empacotar” as aplicações web em arquivos executáveis nos próprios computadores dos usuários. Isso poderá ser útil para casos em que o usuário não consiga acesso à intranet do Centro.

4. Outras atividades

Padronização de processos de análise e de desenvolvimento de protótipos

Para um gerenciamento efetivo das múltiplas demandas de tipos de análises de dados do Centro, foram propostas atividades para formalizar e estruturar os processos de criação conceitual de ferramentas com foco na prototipagem rápida das ideias e algoritmos discutidos pela equipe do projeto.

O trabalho foi iniciado com um levantamento de bases de dados e algoritmos existentes no CGEE da perspectiva dos usuários, incluindo suas descrições e exemplos de uso. Essa iniciativa, reportada no produto de 2020, foi pensada para organizar, sob a ótica das assessorias e equipes de projetos, os recursos de informação do Centro. As referências de bases de dados e ferramentas existentes foram obtidas da equipe de TI e estruturadas sob a forma de um organograma navegável em ferramenta especializada para esse tipo de visualização. Como esse acervo é dinâmico, em um momento conveniente para as equipes envolvidas, sugere-se a entrega do organograma final no formato HTML e hospedá-lo em uma página na intranet do Centro.

Em uma segunda etapa, foi elaborada, ainda em 2020, uma lista de melhores práticas, juntamente com testes e adoção das alternativas testadas, para desenvolvimento de protótipos com base em padrões de processamento de linguagem natural e criação de aplicativos executados juntamente com interfaces web. Com base nessa atividade a equipe elaborou um fluxo de trabalho que contempla a escolha de algoritmos, suas validações como *notebooks* do Jupyter, suas implementações em linha de comando da linguagem Python, escolhida pelo já bastante expressivo acervo de bibliotecas de soluções e algoritmos desenhados para análises de dados, além da integração dos códigos validados com interfaces web que facilitam o uso e testes por parte de usuários das assessorias.

Na fase final desse processo foi realizada a consolidação da lista de melhores práticas elaborada em 2020 a práticas incorporadas pela equipe do projeto ao longo de 2021, sob a forma de um documento contendo uma proposta de padrões para desenvolvimento de protótipos de ferramentas de análise de dados para o CGEE.

Discussão contínua sobre conceitos e técnicas inovadores

Com o objetivo de subsidiar a formulação de estratégias futuras de atuação do CGEE na análise e visualização de dados, informação e conhecimento, foi incorporada às metas do projeto EDVI uma atividade que prevê o estudo contínuo de conceitos e técnicas inovadores em Ciência de Dados e Representação do Conhecimento com impacto estratégico para o Centro.

Para a realização dessa meta, foi criado um ambiente de discussão institucional virtual apropriado às atividades previstas. Condizente com a natureza experimental da ciência de dados visando a exploração de ideias no longo prazo, foram realizadas discussões conceituais, explorações de novas bases de dados, testes de novos algoritmos e bibliotecas, exposições de ideias inovadoras e o desenvolvimento de protótipos, vários deles descritos nas seções anteriores. Em todos os casos, as discussões tiveram algumas diretrizes básicas.

Como primeira diretriz, as discussões sempre buscaram promover a troca de conhecimentos técnicos e experiências entre os membros do grupo, seja no que diz respeito a processos de desenvolvimento, seja na fundamentação matemática dos algoritmos de *back end*, seja na fundamentação conceitual de soluções de *front end*. Como consequência lógica dessa diretriz, as discussões tenderam a focar mais em algoritmos, técnicas de otimização de códigos e na fundamentação de conceitos de design, ergonomia e experiência do usuário final do que em ferramentas e bibliotecas.

Uma outra diretriz, compatível com a anterior, foi a de delimitar a concepção de futuras ferramentas com algoritmos originais que tenham potencial de prover saltos qualitativos à capacidade já existente do Centro. A visão de futuro desses protótipos de ferramenta não visava usualmente competir com soluções existentes no mercado, mas também não excluia a possibilidade de testá-las para descartá-las, simplesmente usá-las ou adaptá-las a objetivos específicos. Condizente a essa diretriz, as soluções discutidas foram quase que exclusivamente de código aberto.

A última diretriz foi a busca ativa de demandas de outros projetos do CGEE para ajudar na

concepção de novas soluções. Essa tarefa foi facilitada pela forma matricial de atuação dos empregados do CGEE, particularmente dos hoje 7 membros do grupo de discussão que, em conjunto, colaboraram com cerca de uma dúzia projetos diferentes do Centro. Dentre os conceitos, soluções de software e métodos discutidos, destacam-se:

a) Testes e aplicações de ferramentas para processamento de grandes volumes de dados – Essa atividade iniciou como resposta a uma demanda da diretoria do CGEE que envolvia processar centenas de milhares de registros das bases de dados de artigos do WoS do OCTI para estimar a evasão de pesquisadores brasileiros para o exterior. Para essa tarefa foram testadas algumas ferramentas de código aberto para o tratamento de grandes volumes de dados (*big data*). Como *framework* para armazenamento foi empregado o Apache Spark, sistema considerado por muitos como alternativa melhorada ao MapReduce como *engine* unificador de técnicas analíticas de processamento de dados em larga escala. Um outro teste associado foi o emprego da linguagem de programação Scala na codificação das soluções propostas. Trata-se de uma linguagem compatível com Java que tem sido consistentemente usada para substituí-lo em sistemas mais modernos. Os primeiros resultados foram promissores e existem possíveis aplicações futuras que poderão se beneficiar desse aprendizado.

b) Realização de estudos, treinamentos e testes de uso das ferramentas de *big data* da *Amazon Web Services* (AWS) – Tratou-se de uma demanda do MCTI no contexto do projeto "Formatos e práticas inovadoras para o financiamento do SNCTI", onde se buscou classificar páginas com ofertas de financiamento previamente catalogadas por equipes do Ministério como de possível interesse. A demanda consistia em usar as páginas pré-classificadas para treinar algoritmos de *machine learning* e compará-los com as ferramentas disponíveis no *Amazon Comprehend*, serviço de processamento de linguagem natural do pacote de serviços de nuvem AWS. Essa atividade foi muito importante para a equipe do projeto ter um primeiro contato com serviços de nuvem, suas potencialidades e limitações.

c) Estudo de otimização de desempenho de algoritmos de *machine learning* – As maiores limitações observadas no uso da AWS foram relacionadas aos altos preços cobrados pelo acesso aos seus serviços de inteligência artificial. Para o objetivo desejado, a classificação automática de páginas com fontes de financiamento de potencial interesse para pesquisadores brasileiros, essa limitação foi amplificada. Ocorre que o conjunto de páginas

pré-classificadas pelo MCTI era muito pequeno para termos um treinamento adequado dos modelos de aprendizado de máquina, o que dificultou uma avaliação adequada de performance dos algoritmos da AWS para a tarefa desejada.

O desempenho de modelos é medido através de métricas que são fortemente dependentes das flutuações estatísticas inerentes à seleção aleatória dos subconjuntos de treinamento, validação e teste. Para minimizar artefatos estatísticos nos valores dessas métricas, o padrão que se usa é repetir 10 vezes o mesmo experimento gerando 10 modelos com os mesmos parâmetros dos algoritmos, mas embralhando os elementos dos conjuntos de treinamento, validação e teste. Como, para conjuntos de dados pequenos como o do Ministério, as flutuações estatísticas eram muito altas, seriam necessários mais de 10 desses experimentos a um custo desconhecido, caso o número mínimo para um nível de confiança considerável razoável para a estimativa das métricas fosse alcançado. Um extenso estudo de validação cruzada foi realizado nos computadores do CGEE (sem custo) e concluiu-se que os resultados de desempenho mais confiáveis seraim alcançados com cerca de 40 experimentos, o que deu aos gestores do projeto a possibilidade de otimizar o uso do *Amazon Comprehend* com confiança mínima nas métricas de desempenho. Um exemplo apenas ilustrativo para uma das baterias de testes para as métricas de desempenho pode ser visto na Figura 11 abaixo.

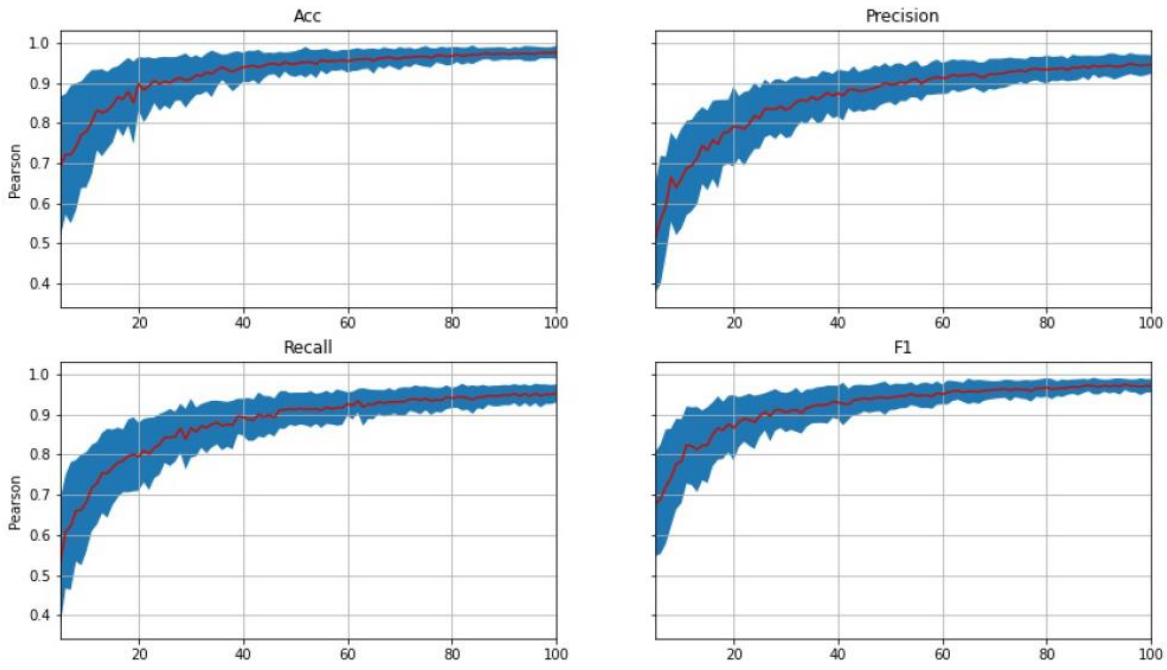


Figura 10: Exemplo de bateria de testes para a otimização do número de experimentos de validação cruzada necessários para estimar a confiabilidade das métricas de desempenho de modelos de aprendizado de máquina aplicados a textos (mais detalhes no texto acima). Este exemplo se refere a uma rede neural do tipo LSTM.

d) Estudo sobre usos de modelagem de tópicos com base em alocação latente de Dirichlet – Como dito anteriormente, o CGEE emprega há anos métodos de análise de redes para processar textos, o que chamamos de redes de similaridade semântica entre documentos. Ocorre que nos últimos anos tem havido um grande interesse na literatura especializada por técnicas estatísticas para a modelagem de tópicos. A mais popular dessas técnicas é a alocação latente de Dirichlet. O objetivo do estudo foi avaliar vantagens comparativas com relação às técnicas já empregadas. Foi concluído que eventuais soluções baseadas em alocação latente de Dirichlet têm que ser validadas caso a caso e que os algoritmos desenvolvidos eram suficientes para o uso *ad hoc*, antes de se buscar a implementação de um protótipo específico, com seu respectivo custo de recursos do Centro. Um exemplo de agrupamento de documentos usando a técnica t-SNE é mostrado na figura abaixo:

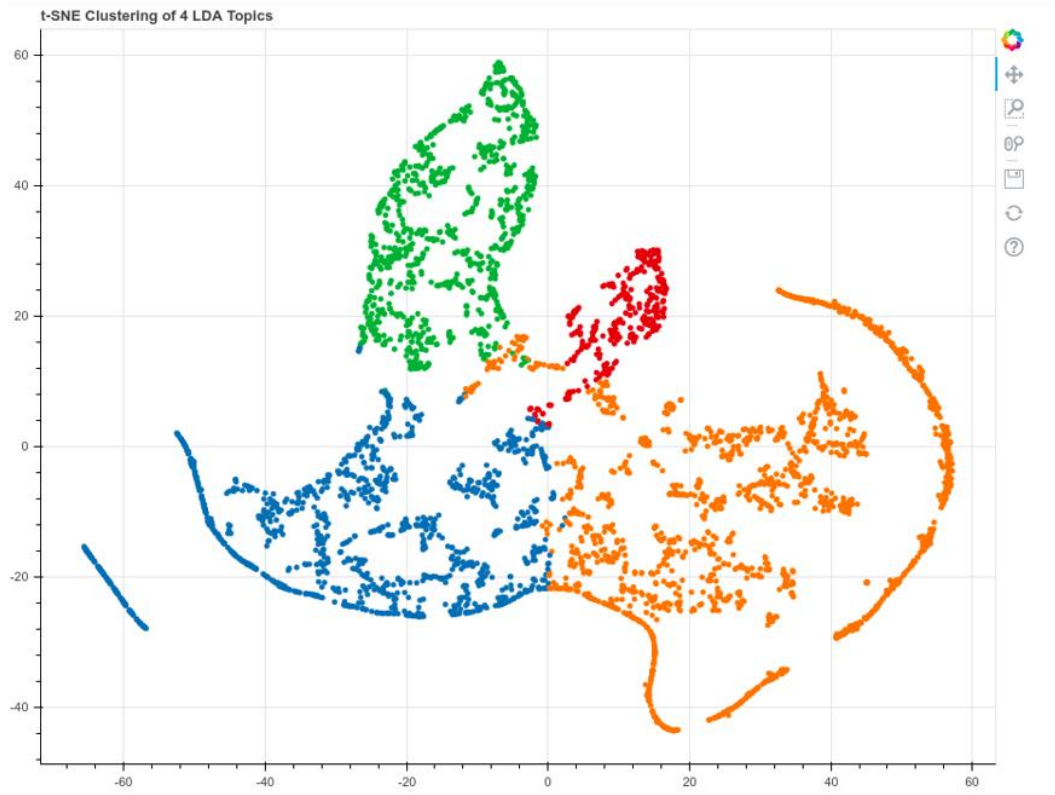


Figura 11: Representação de agrupamentos de documentos onde a modelagem de tópicos foi realizada com base em alocação latente de Dirichlet. Cada tópico corresponde a uma cor.

e) Otimização de modelagem com redes neurais com operadores neurais de Fourier, com possíveis aplicações na identificação de sinais de temas emergentes em dados de interesse do Centro. Foi avaliado que a técnica, embora potencialmente interessante para uso no Centro, compete com soluções mais simples que já temos à mão.

Desenvolvimento de metodologias relacionadas à inovação e à proteção de propriedade intelectual

Para consolidar no CGEE metodogias para análises de dados de patentes, um objetivo específico nessa direção foi estabelecido para o projeto EDVI em 2021. Ao longo do ano, tivemos várias discussões sobre diversas atividades relacionadas a essa meta. Destacamos abaixo apenas as mais relevantes:

a) O trabalho iniciou com uma revisão da literatura acadêmica sobre o estado arte em algoritmos para as análises de dados de propriedade intelectual.

b) Parte da equipe do projeto fez cursos introdutórios da Organização Mundial da Proteção da Propriedade Intelectual, para disseminação de conceitos e aprendizado no tratamento de dados de propriedade industrial.

c) Foram realizados testes iniciais de construção e evolução temporal de redes de coocorrência de códigos de tecnologias a partir de dados da base de patentes Derwent.

d) Foi desenvolvimento o piloto de uma nova metodologia para mapeamento temático de redes de tecnologias (com os nós sendo os códigos IPC ou CPC e as arestas, suas coocorrências em registros de patentes). A metodologia foi validada preliminarmente e entrou como contribuição para o boletim sobre economia do Hidrogênio, do projeto "Agenda positiva da mudança do clima e do desenvolvimento sustentável".

e) Foi constituída e testada, para uma avaliação de impacto de leis de incentivo à inovação, uma base de dados de patentes nacionais extraídas da publicação Revista da Propriedade Industrial, do INPI.

f) Foi criado um grupo de discussão com representantes de diferentes projetos do Centro com interesse em propriedade industrial para a harmonização e priorização de interesses sobre análises de dados de patentes. Das reuniões foram estabelecidas prioridades para o desenvolvimento de soluções para análises de dados de patentes em 2022.

Plataforma de acesso aos protótipos e resultados do projeto EDVI

Foi finalizada na intranet do Centro a versão inicial de uma plataforma de acesso aos protótipos desenvolvidos no projeto. Seus conteúdos compreendem a alocação dos códigos dos protótipos desenvolvidos ou em desenvolvimento no gitlab gerenciado pela TI, a definição de um espaço no servidor para acesso por parte dos empregados do CGEE às ferramentas do projeto e a elaboração de uma página com identidade visual para centralizar conceitos, tutoriais e links para os protótipos, com previsão de atualização e manutenção contínuas. Nessa página, por enquanto, estão as soluções mais maduras dentre os protótipos desenvolvidos. O ambiente é orientado a metodologias relevantes para os trabalhos do Centro. Algumas imagens da página podem ser vistas na Figura 12:

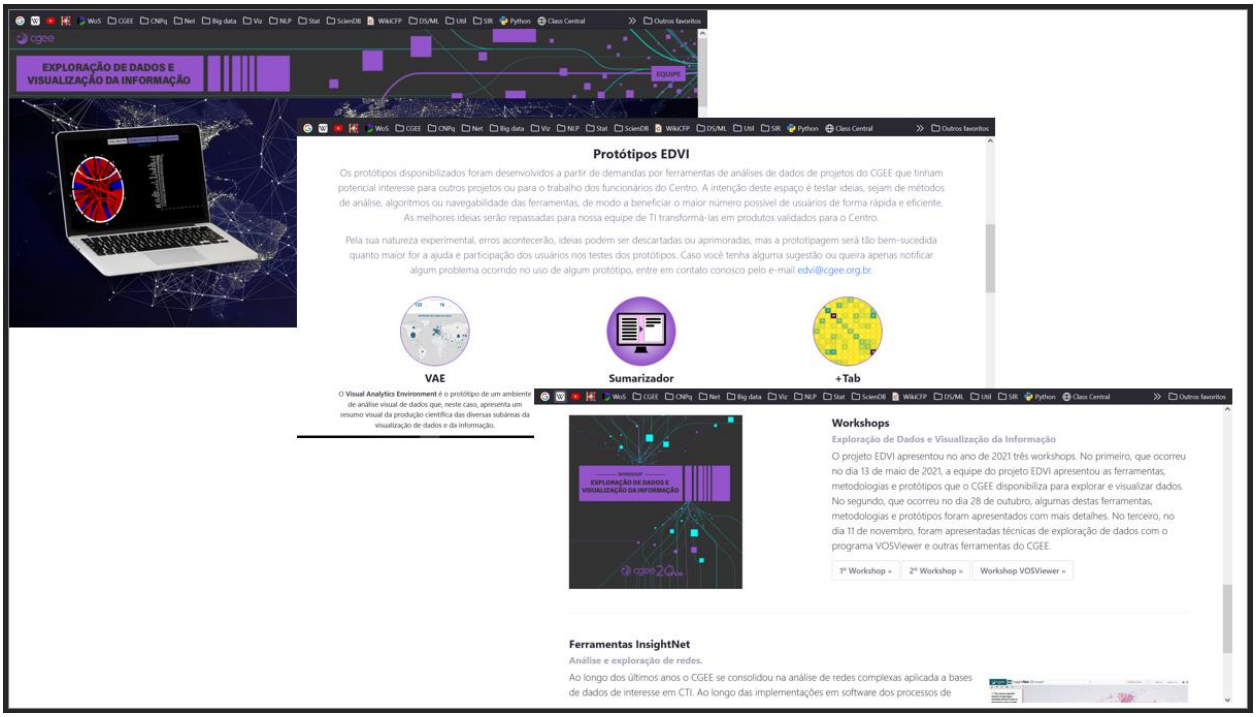


Figura 12: Página de acesso aos protótipos e produtos do projeto EDVI.

Apêndice A: Manual CGEE Insight Net 3.2.10

Sumário

CAPÍTULO 1

Introdução

1.1 Contexto e Visão Geral

O *plugin* **CGEE insight Net** foi concebido para operar junto com o software *Gephi*[1] para viabilizar a análise de grandes volumes de dados disponíveis para o CGEE de modo a organizá-los como redes complexas manipuláveis por usuários com pouca experiência ou treinamento em programação. Essa ferramenta vem sendo continuamente desenvolvida no CGEE desde 2013. A aplicação tem se mostrado eficaz para visualizar redes de coautorias e de similaridade temática entre currículos disponibilizados na Plataforma Lattes do CNPq e de similaridade temática entre artigos disponibilizados em grandes bases de dados como *Scopus* e *Web of Science*, embora outras fontes textuais também possam ser ex- ploradas. Este guia de usuário descreve a versão 3 do *plug-in*, com detalhes sobre a instalação dos seus componentes e suas principais funcionalidades.

[1] http://www.gephi.org

1.2 Ajuda online

Este manual de usuário está disponível online no **CGEE Insight Net**. No menu *Help* existe a opção

CGEE Insight Net Help:

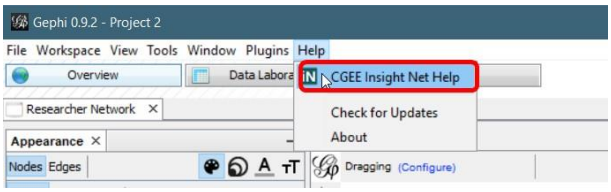


Figura 1.1: Opção para abrir a ajuda on-line

1.3 Funcionalidades experimentais

O CGEE Insight Net está em um processo de aprimoramento constante e a grande maioria das funciona- lidades se encontra em um estado robusto e estável. Outros métodos e algoritmos foram acrescentados apenas recentemente e podem apresentar instabilidades ou deficiências no processamento de dados es- pecíficos. Essas funcionalidades estão marcadas no *plugin* com a plavra `EXPERIMENTAL` ou o símbolo ⚠.

1.4 Envio do protocolo de execução

Em caso de erros inesperados no **CGEE Insight Net**, o usuário pode enviar um relatório de erros ao CGEE. Quando acontecer um erro inesperado, a seguinte mensagem é exibida:

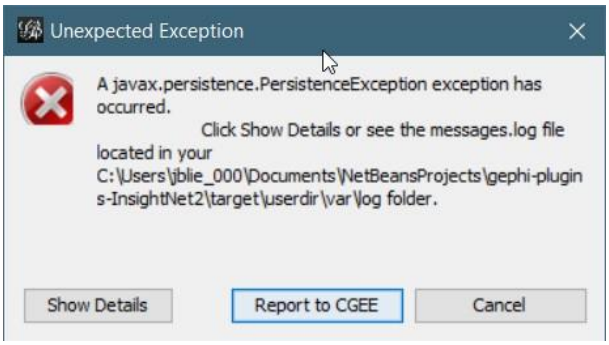


Figura 1.2: Mensagem de erros inesperados

Clicando em *Report to CGEE*, o **CGEE Insight Net** mostra o seguinte

diálogo:

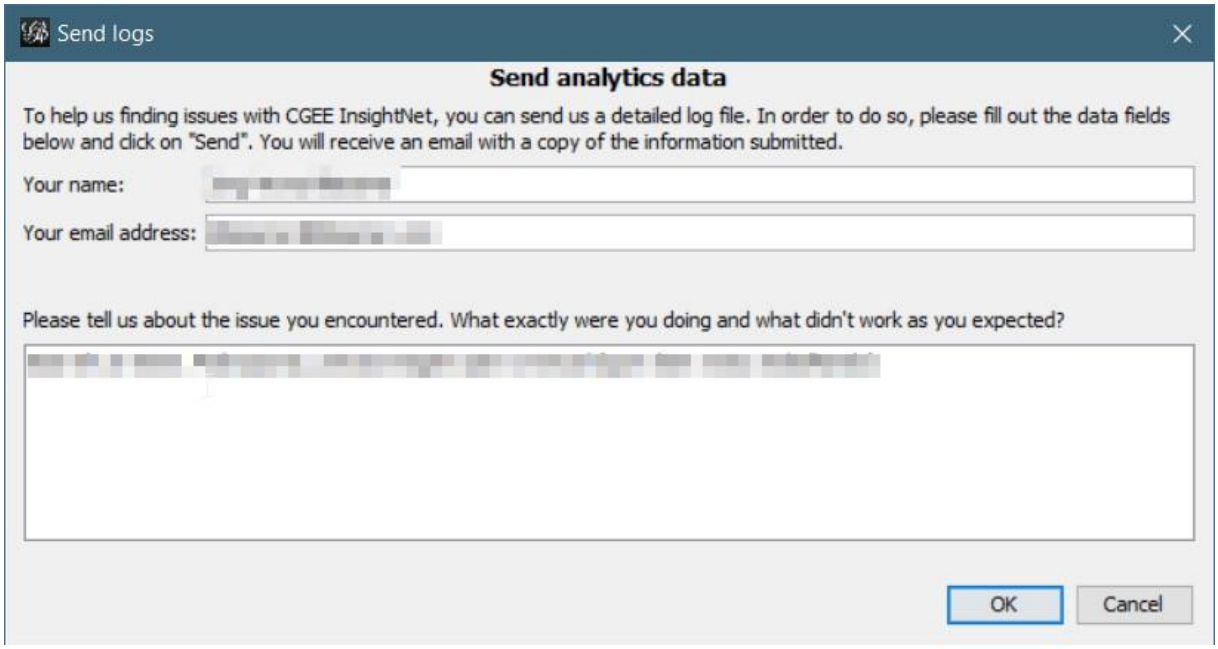


Figura 1.3: Diálogo de envio do protocolo de execução

Recomenda-se preencher todos os campos desse diálogo para que o CGEE possa reproduzir e eliminar possíveis problemas. Depois, o usuário deve clicar em "**OK**" e os dados são enviados. No final deste processo, o resultado do envido é exibido:

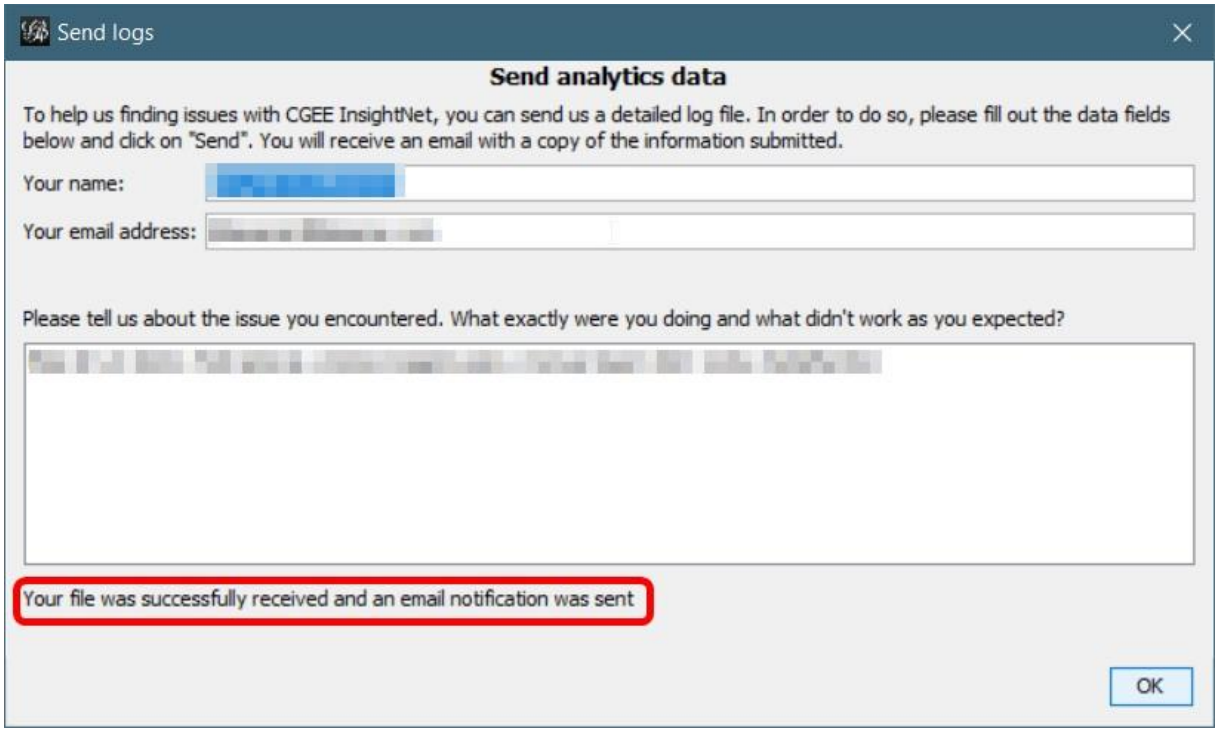


Figura 1.4: Conclusão do envio do protocolo de execução

O **CGEE Insight Net** também detecta se o *Gephi* for encerrado de forma irregular. Neste caso, a seguinte mensagem é exibida quando o programa for reiniciado:

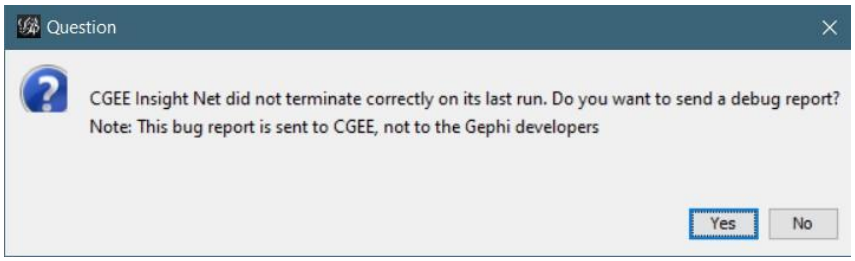


Figura 1.5: Mensagem após o encerramento irregular do Gephi

Clicando em "*Yes*", o diálogo de envio do protocolo de execução é exibido e deve ser preenchido con- forme descrito anteriormente.

CAPÍTULO 2

Instalação do CGEE Insight Net

2.1 Pré-requisitos

O *CGEE Insight Net* usa a versão 0.9.2 do software Gephi. Na data da atualização do presente manual de sistemas. A versão 0.9.2 do Gephi depende da instalação do ambiente Java na versão 1.8 nos ambientes de Windows, Linux e macOS.

2.2 Instalação do software Gephi

O software Gephi pode ser baixado na versão 0.9.2 pela página da ferramenta na internet:

https://github.com/gephi/gephi/releases

O software vem com um instalador automático que suporta os principais sistemas operacionais. O processo é documentado no site do Gephi[1]

As principais características do Gephi podem ser revisadas na página https://gephi.org/features/.

Caso sejam instaladas diversas versões do Java no computador do usuário, o Gephi permite a configura- ção de qual delas deve ser usada. Para isso, existe uma variável `jdkhome` no arquivo `gephi.conf` que consta no subdiretório `etc` da instalação do Gephi. **Recomenda-se tornar este arquivo ``gephi.conf`` gravável para o usuário comum**.

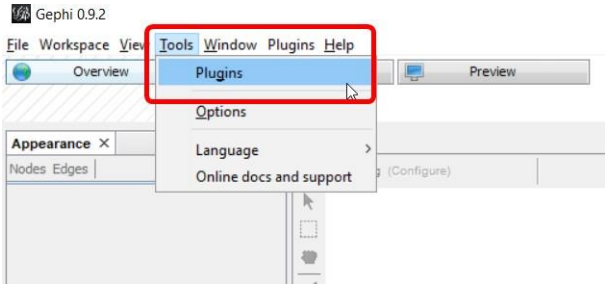
[1] https://gephi.org

2.3 Configuração da central de atualizações

Esse passo é necessário apenas uma vez para cada computador. O *CGEE Insight Net* é disponibilizado online, no seguinte endereço:

http://analise-rede.pages.cgee.org.br/insightnet-plugin/updates.xml

Para instalar o módulo, esse endereço deve ser especificado na tela *Tools* > *Plugins*, na aba *Settings* do Gephi, clicando no botão "*Add*":



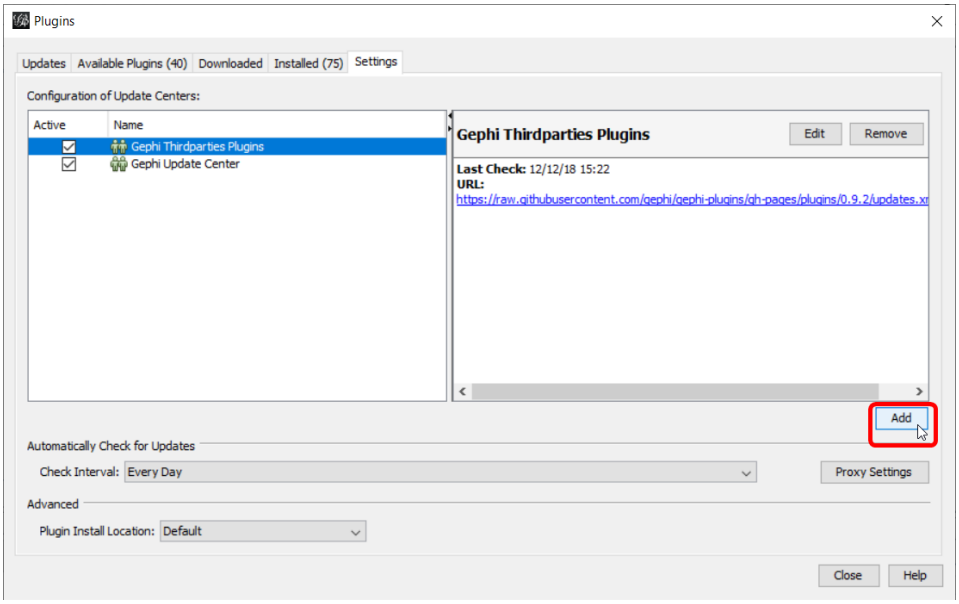


Figura 2.1: Configuração dos centrais de

atualização Gephi No diálogo que aparece, os seguintes dados

devem ser especificados:

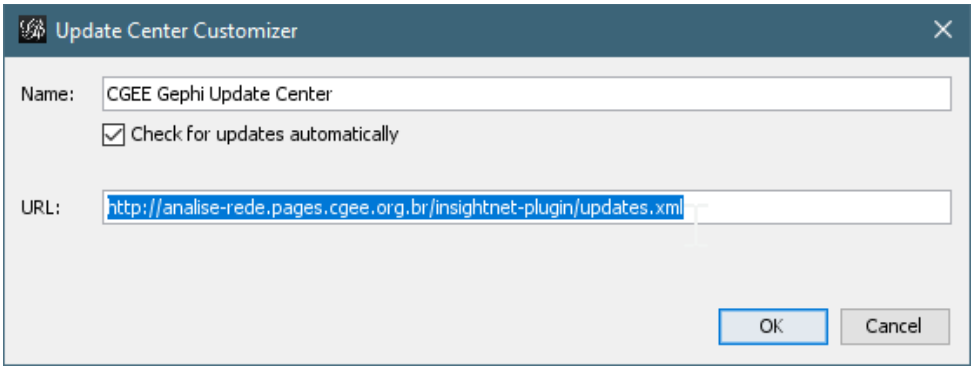


Figura 2.2: Configuração da central de atualizações do CGEE Insight Net

Com isso, a central de atualização aparecerá na lista de configuração. Caso seja necessário, o usuário deve configurar o *proxy* da conexão com a internet (botão *Proxy Settings*).

2.4 Instalação do *CGEE Insight Net*

Com a central de atualização configurada, o usuário pode selecionar e instalar o *CGEE Insight Net*, também pela tela de *plug-ins* (*Tool* > *Plugins*). Clicando na aba "*Available Plugins*", a ferramenta mostra uma lista de todos os *plug-ins* disponíveis, entre eles o "*CGEE Analysis plugin*", que deve ser selecionado pelo usuário, seguido por um clique no botão "*Install*":

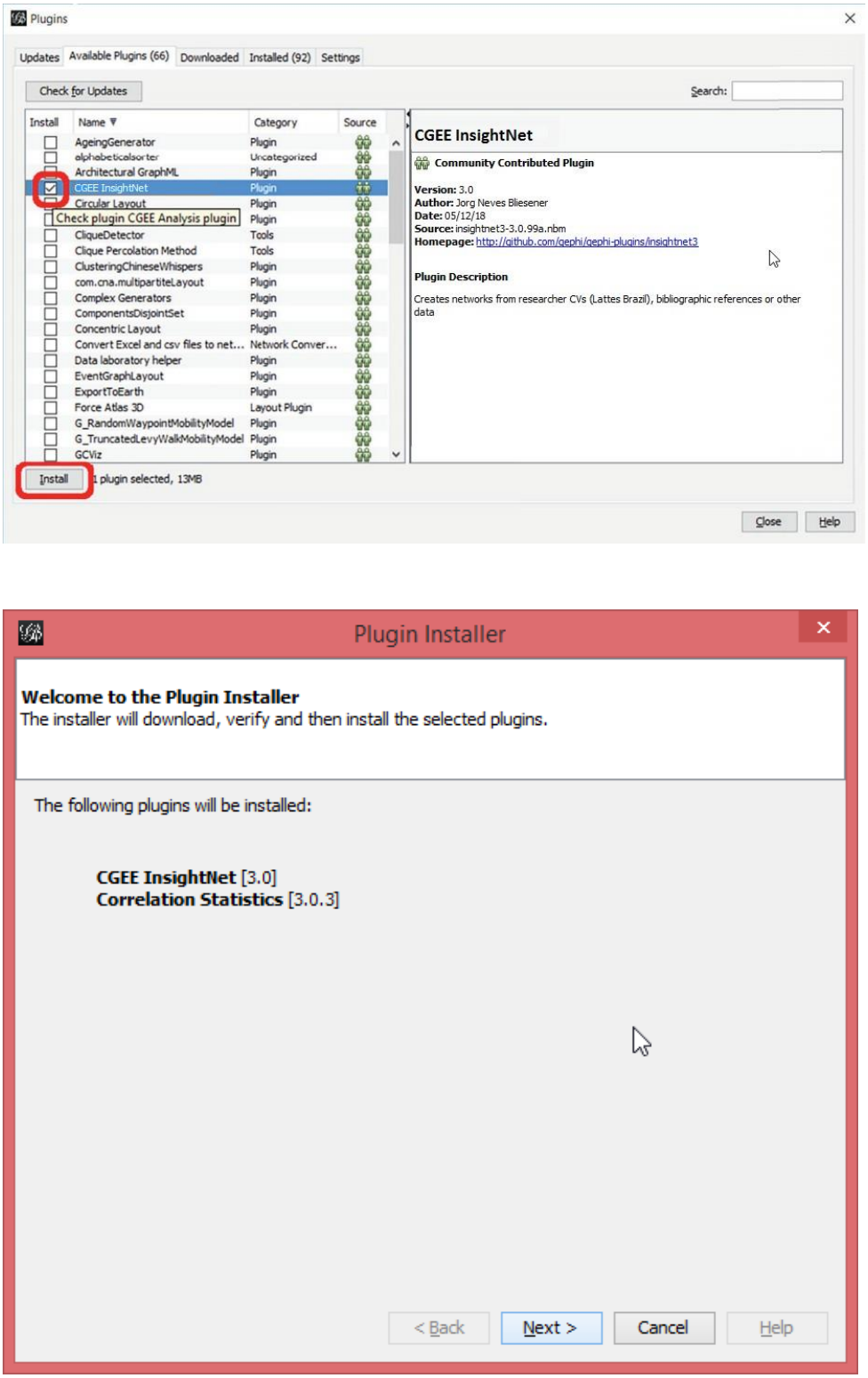


Figura 2.3: Instalação do CGEE Insight Net

Ao chegar à tela de licença, o usuário deve aceitar o texto da(s) licenças exibida(s) para concluir a instalação[2] :

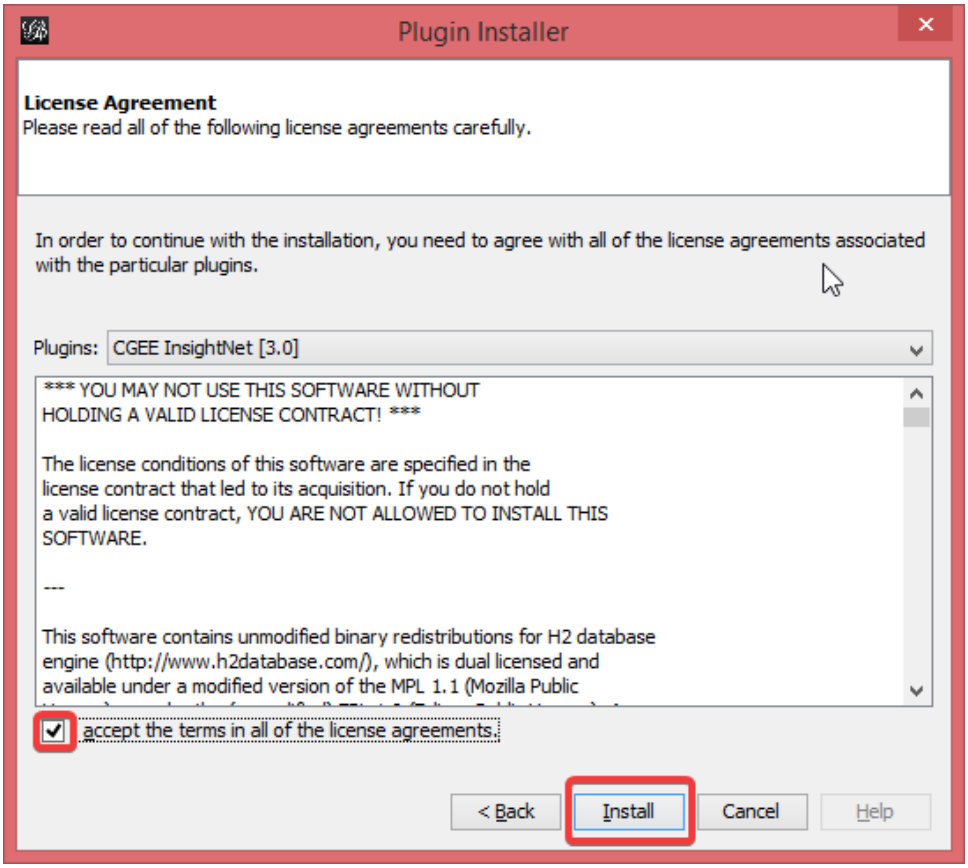
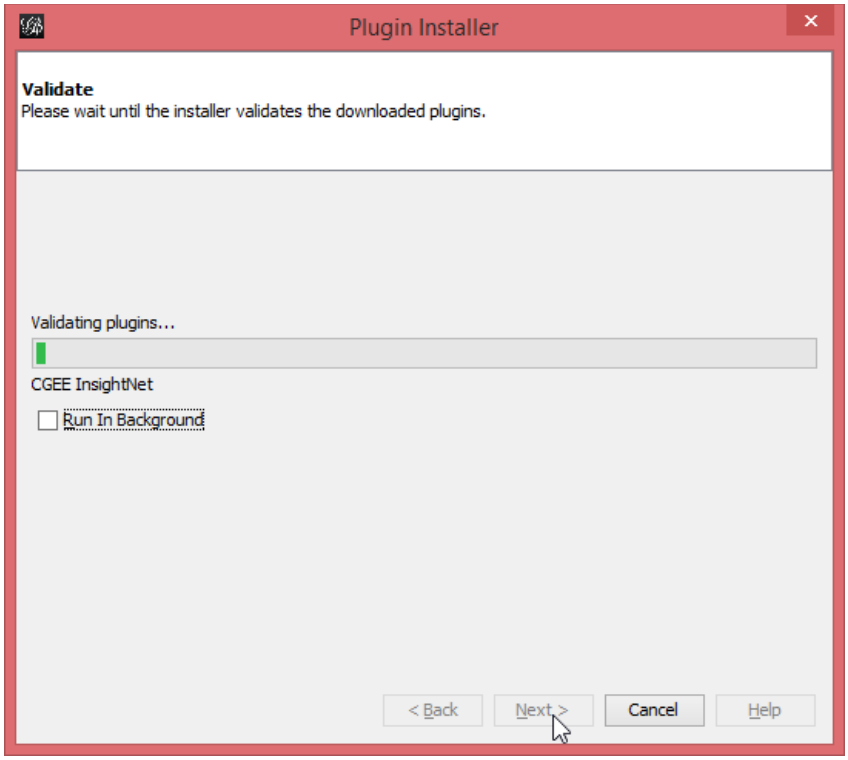


Figura 2.4: Concordância com a(s) licenças(s) do produto

Depois disso, o *CGEE Insight Net* é baixado pela internet e um aviso de falta de assinatura digital é exibido, e pode ser ignorado:

[2] As licenças exibidas incluem as licenças das bibliotecas usadas no produto.



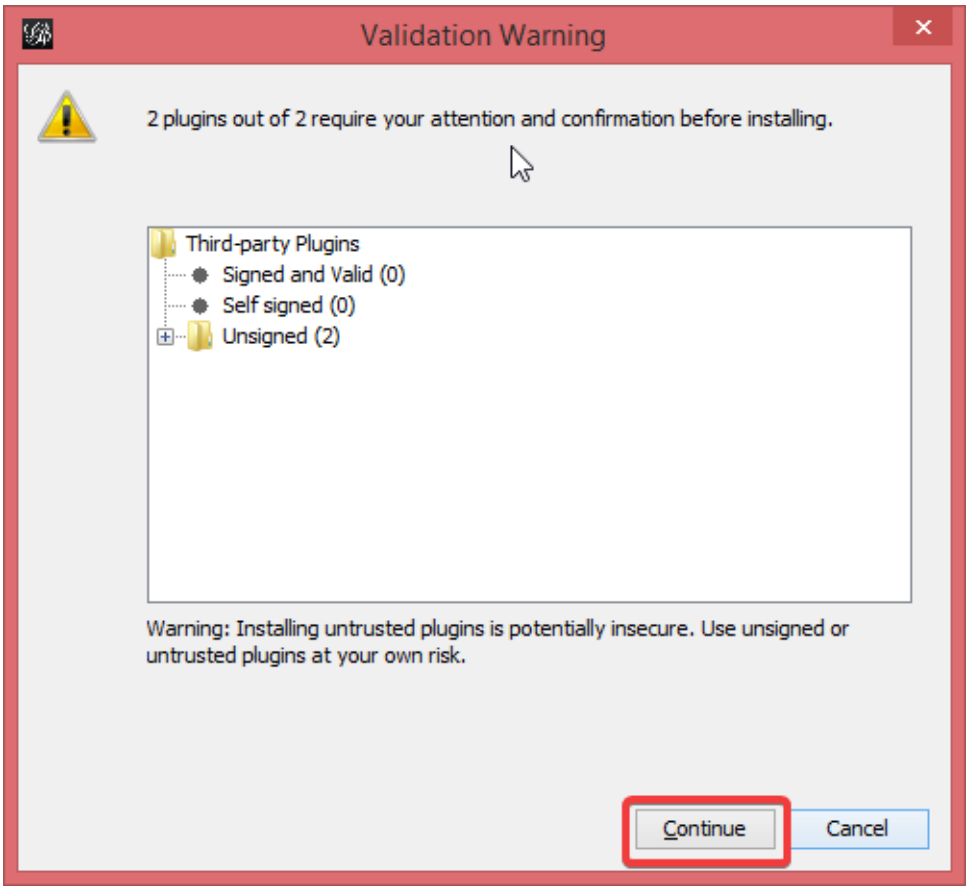


Figura 2.5: Download e aviso de falta de assinatura

Em seguida, o Gephi **deve ser reiniciado e a instalação do** CGEE Insight Net **é concluída**:

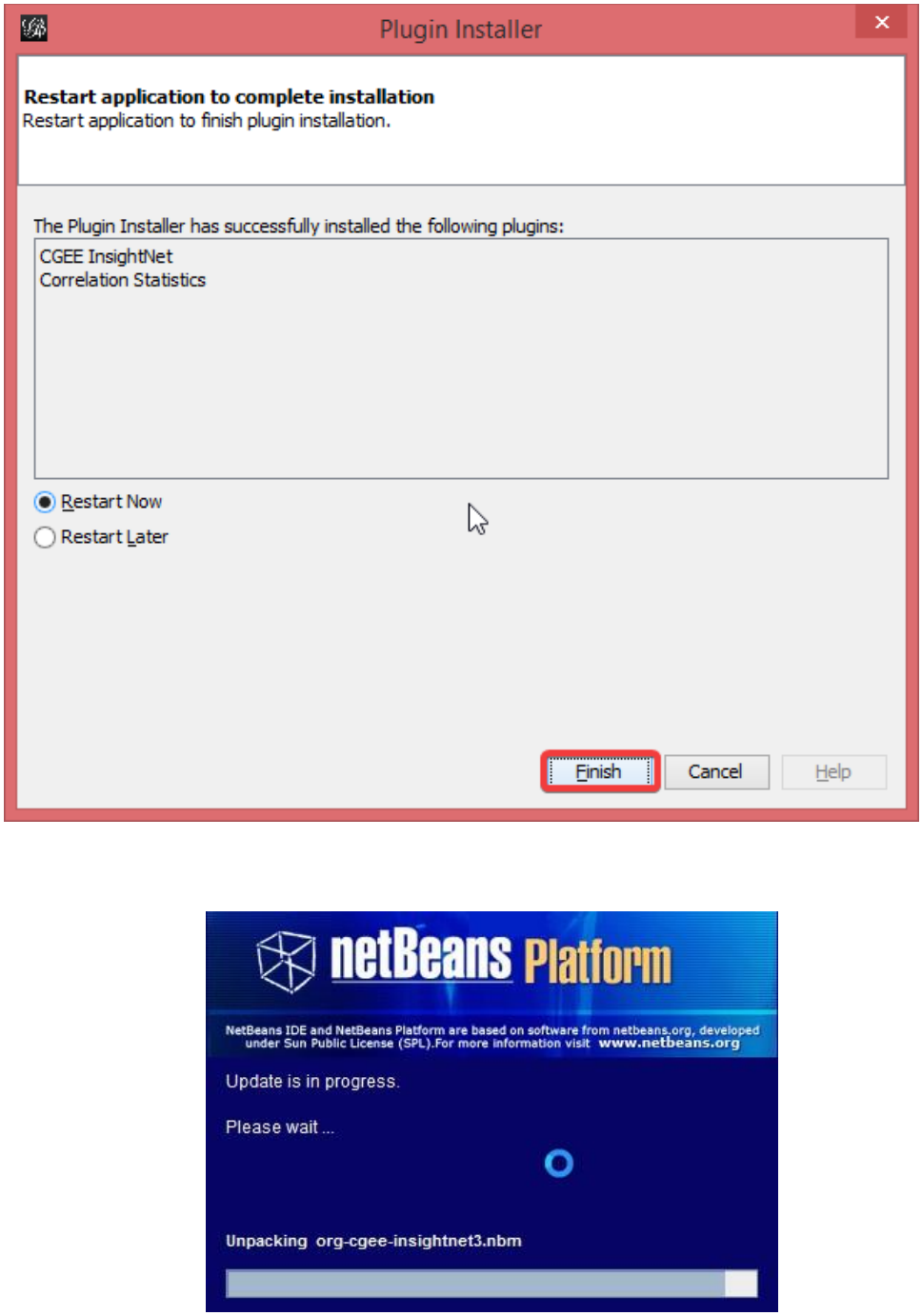


Figura 2.6: Instalação do CGEE Insight Net

Depois de reiniciar, o *GEE Insight Net* tenta alterar o arquivo `gephi.conf` que consta no subdiretório `etc` da instalação do Gephi[3]. Caso essa tentativa tenha êxito, a seguinte mensagem é exibida e o sistema é reiniciado outra vez:

[3] Conforme relatado na Seção 2.2, recomenda-se que o administrador do Sistema torne esse arquivo gravável para o usuário

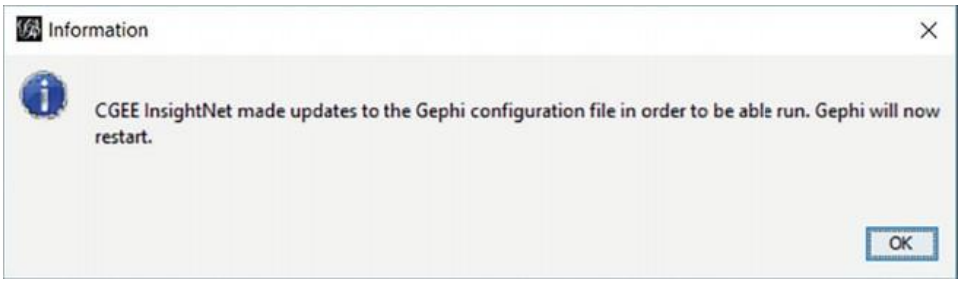


Figura 2.7: Mensagem de alteração bem-sucedida do arquivo gephi.conf

Caso o arquivo `gephi.conf` não possa ser alterado, a seguinte mensagem é exibida:

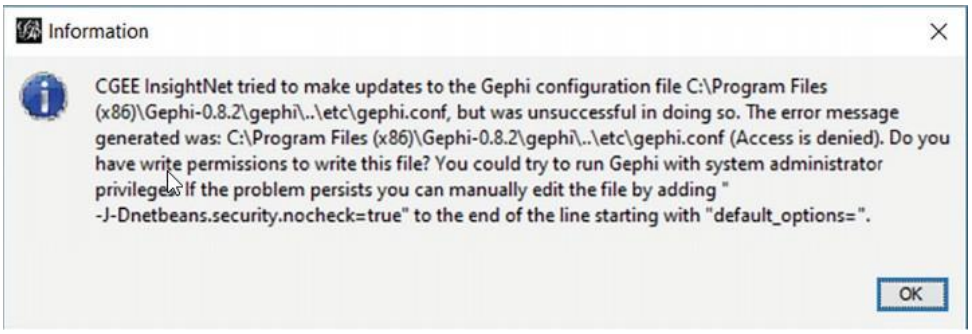
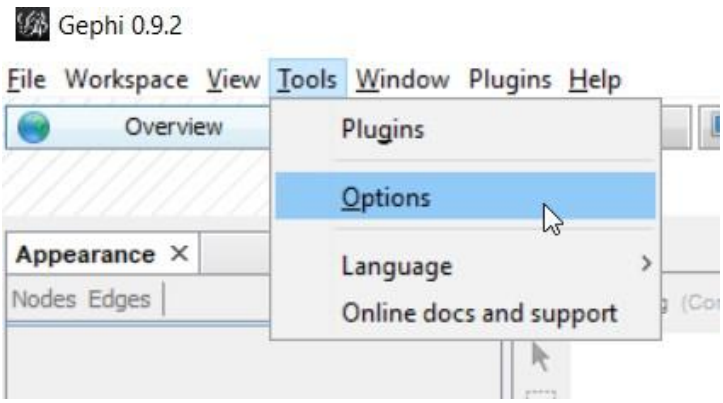


Figura 2.8: Mensagem de alteração malsucedida do arquivo gephi.conf

Nesse caso, a gestão e o uso de licenças adicionais (ver Seção 3.7) fica indisponível e uma das alte- rações sugeridas no diálogo deve ser realizada por um administrador do sistema. A solução de menor complexidade é a concessão dos privilégios de gravação do arquivo `gephi.conf` para o usuário final, conforme descrito na Seção 2.2.

Depois desse procedimento, o *CGEE Insight Net* aparece na tela de opções do Gephi. Para usar efeti- vamente, as licenças correspondentes aos módulos do contratados ainda devem ser instaladas conforme descrito na Seção 3.7.



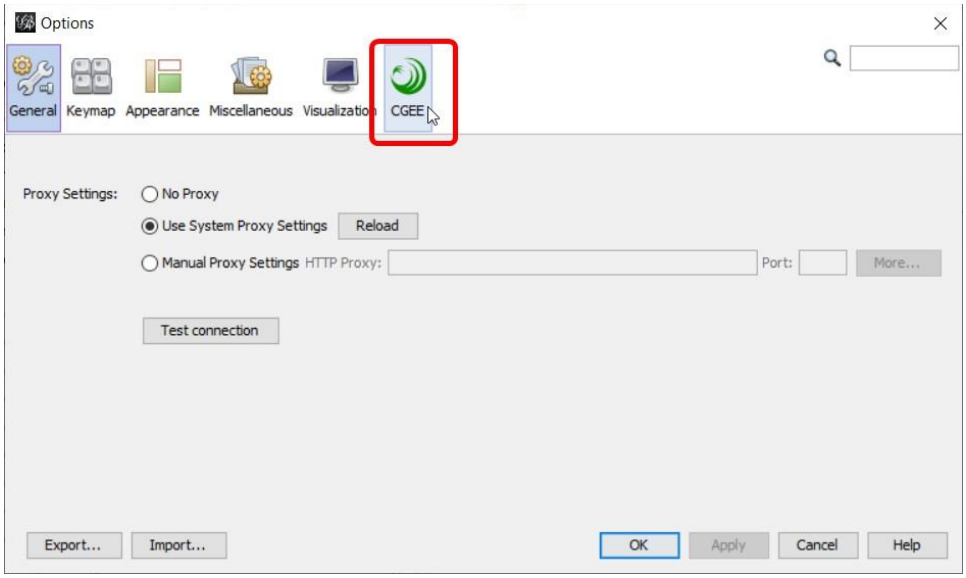


Figura 2.9: CGEE Insight net instalado

2.5 Atualização do *CGEE Insight Net*

O *CGEE Insight Net* é atualizado automaticamente, de acordo com a configuração de atualizações na configuração dos *plug-ins*:

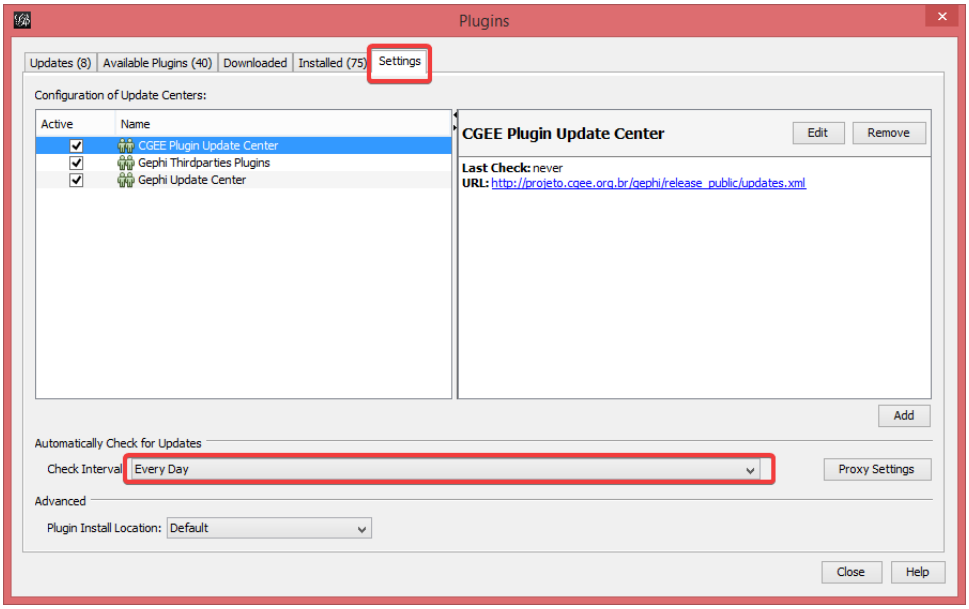


Figura 2.10: Configuração da atualização

automática Caso haja atualizações do *CGEE Insight Net*, um aviso

aparecerá no Gephi:

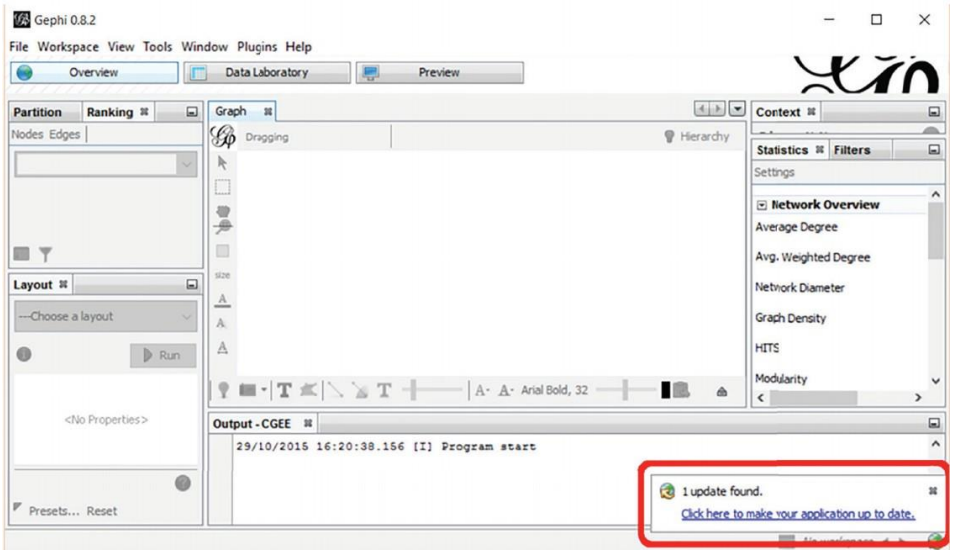
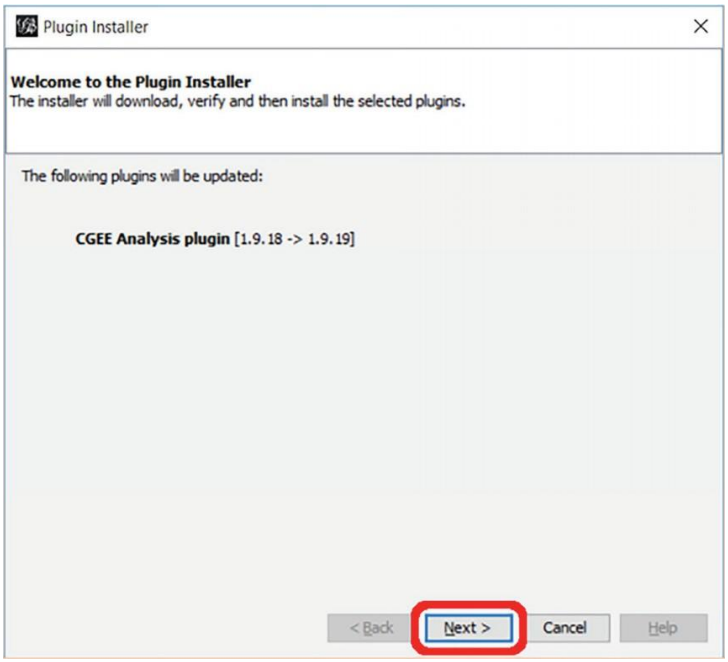


Figura 2.11: Aviso de atualização

Clicando na mensagem de atualização, o Gephi exibe a lista de atualizações disponíveis e, clicando em

Next, inicia o processo de atualização:



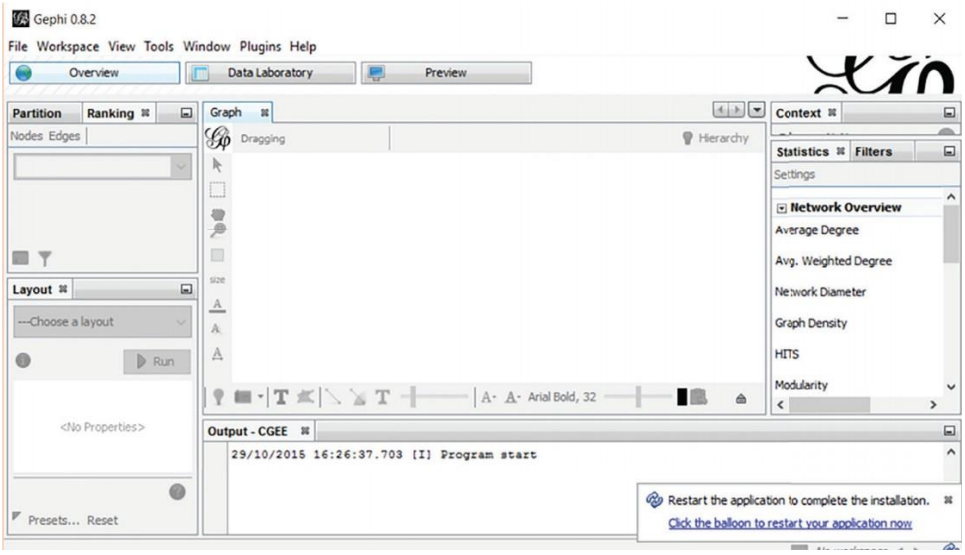


Figura 2.12: Aviso de

atualização Clicando no aviso, o Gephi conclui a atualização e

reinicia o programa.

CAPÍTULO 3

Configuração do CGEE Insight Net

Antes de usar o *CGEE Insight Net*, este deve ser configurado de acordo com os requisitos do usuário, clicando em *Tools* > *Options* e selecionando a tela *CGEE*.

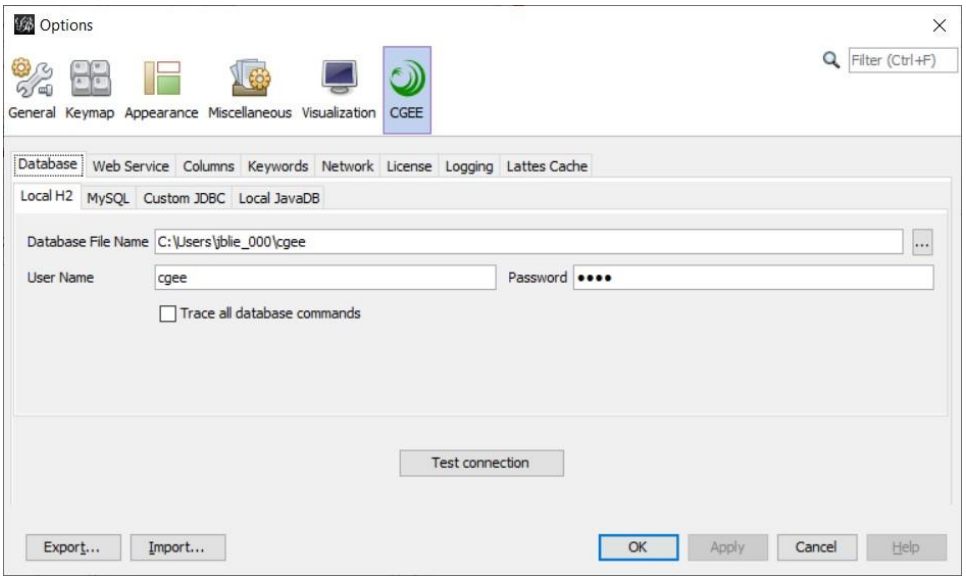


Figura 3.1: Opções de configuração

No caso mais simples de uso local do *CGEE Insight Net*, o usuário apenas confirma as informações pré- configuradas, sem necessidade de alterar qualquer um dos valores. Para

atender casos específicos (banco de dados centralizados, acompanhamento detalhado ou depuração de problemas, redução da carga do computador, etc.), os itens de configuração serão explicados em seguida.

3.1 Configuração do banco de dados

A primeira aba da configuração refere-se ao tipo de banco de dados e aos parâmetros de configuração da conexão.

Para o uso local do *CGEE Insight Net*, recomenda-se o uso do banco "*Local H2*". O nome do arquivo pode ser selecionado pelo usuário. Para conectar, deve ser especificado um nome de usuário e uma senha, sendo que os valores pré-configurados devem atender a maioria dos casos. A customização do usuário e da senha permite certo nível de proteção de acesso, mas não envolve nenhum tipo de criptografia no nível binário da base.

O uso do *CGEE Insight Net* em ambientes centralizados geralmente envolve um banco de dados em outro servidor. No caso do banco "*MySQL*", o caminho do módulo de conexão (o *Driver JDBC*) e os parâmetros de conexão (servidor, porta, usuário, senha e nome do banco de dados) precisam ser definidos na aba correspondente e serão fornecidos pelo administrador do servidor.

Outros bancos de dados podem ser configurados na aba "*Custom JDBC*", o que, geralmente, ainda requer a customização dos comandos SQL envolvidos. Caso necessário, sugere-se o envolvimento de especialistas do departamento de TI.

A aba "*Local JavaDB*" permite o uso de um banco legado, implementado em Java, caso os métodos anteriores não produzam os efeitos desejados. Ressalta-se que o banco JavaDB possui desempenho inferior em relação ao banco H2 sugerido na configuração padrão.

Depois da configuração dos parâmetros do banco de dados, a conexão deve ser verificada, clicando no botão "*Test connection*".

3.2 Configuração do usuário para acessar o banco de dados de Currículos Lattes do CGEE

O acesso ao banco de dados de Currículos Lattes do CGEE é restrito por usuário e senha. A aba "*Web service*" permite a configuração do nome do usuário e da respectiva senha anteriormente criados no *web service* do CGEE:

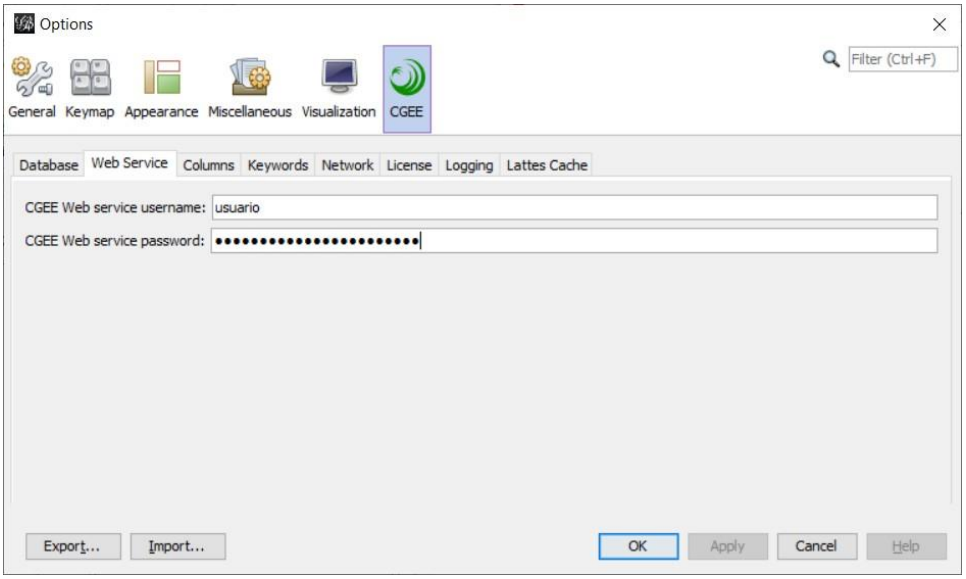


Figura 3.2: Configuração do usuário e da senha para acessar o banco de dados de Currículos Lattes do CGEE

3.2. Configuração do usuário para acessar o banco de dados de Currículos Lattes do CGEE

3.3 Configuração das colunas exibidas

As diversas fontes de dados trazem uma grande quantidade de informações, cuja exibição completa pode tornar a operação do programa ineficiente. A relevância dessas informações depende do projeto específico.

Para não sobrecarregar a tela e permitir a exibição dos dados mais relevantes, o *CGEE Insight Net* permite a seleção de atributos dos pesquisadores e das referências bibliográficas e das colunas corres- pondentes no laboratório de dados. A aba "*Columns*" configura, para cada tipo de dados, as colunas que serão exibidas por padrão, sem a necessidade de personalizações do usuário para cada análise:

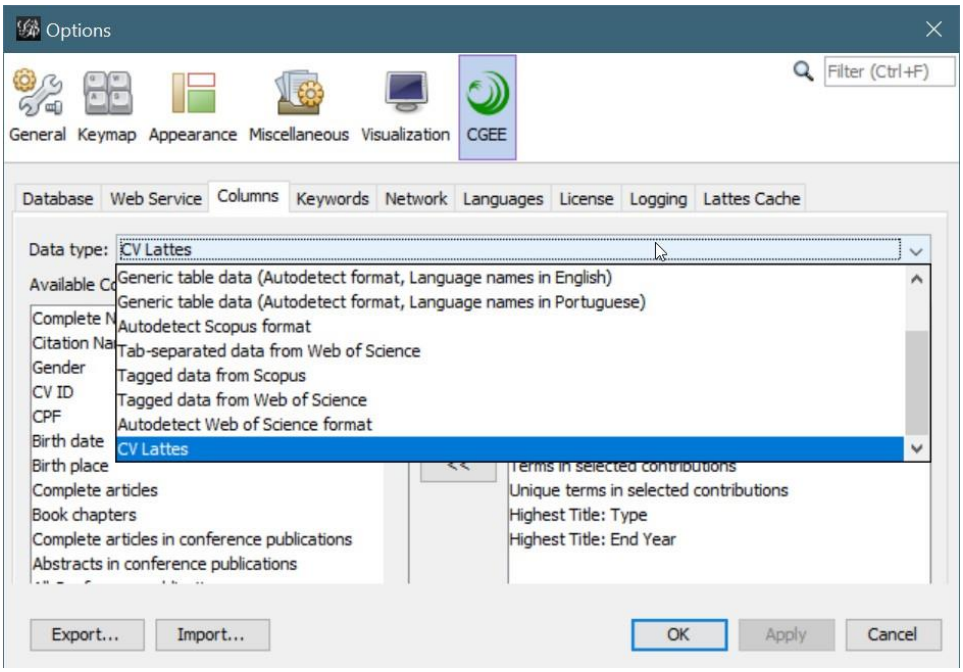


Figura 3.3: Seleção do tipo de dados para configurar as colunas que serão exibidas por padrão

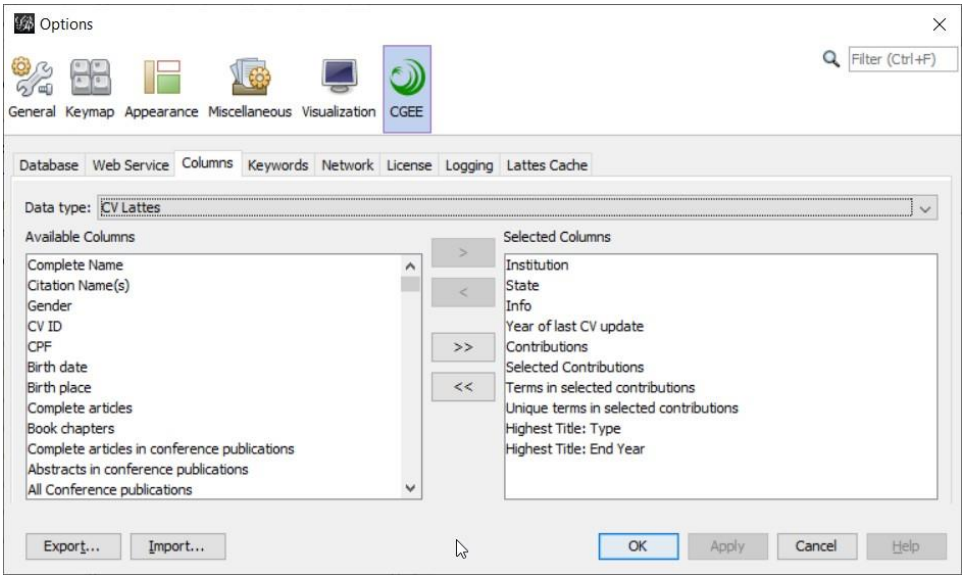


Figura 3.4: Configuração das colunas dos currículos Lattes que serão exibidas por padrão

As colunas na lista da direita são aquelas que aparecem no laboratório de dados. As colunas na lista da esquerda não serão exibidas. O usuário pode clicar em uma ou mais colunas em ambas as janelas segurando a tecla *Ctrl* ou *Shift* e clicar nos botões  ou  para levar essas colunas para a outra lista. Os botões  e  levam todos os elementos de uma lista para outra.

3.4 Exibição da lista de palavras-chave

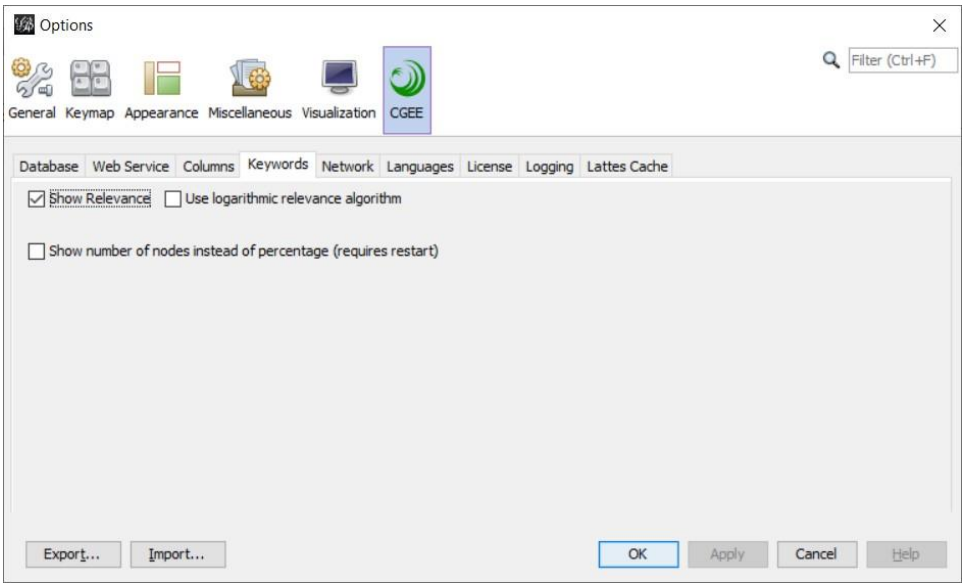


Figura 3.5: Configuração da exibição das palavras-chave

O *CGEE Insight Net* permite a exibição opcional das relevâncias das palavras-chave. Geralmente, a janela de palavras-chave (ver Seção 7.4) mostra, para cada palavra-chave, a sua frequência dentro do conjunto de dados selecionados (pesquisadores ou *clusters*). A opção "*Show Relevance*" permite o uso experimental de um algoritmo que calcula a relevância das palavras-chave a partir do algoritmo "tf. idf".

Caso a opção "*Show relevance*" for selecionada, o algoritmo pode usar pesos lineares ou pesos logaritmi- cos para a frequências das palavras, dependendo da configuração da opção "*Use logarithmic relevance algorithm*".

Na janela de palavras-chave (ver Seção 7.4), a quantidade de nós que referenciam uma palavra-chave pode ser exibida como porcentual da quantidade total de nós ou como número absoluto. A opção "*Show number of nodes instead of percentage*" permite alternar entre essas duas opções. Depois de alterar essa configuração, o *Gephi* precisa ser reiniciado.

3.5 Parâmetros da pesquisa por similaridade

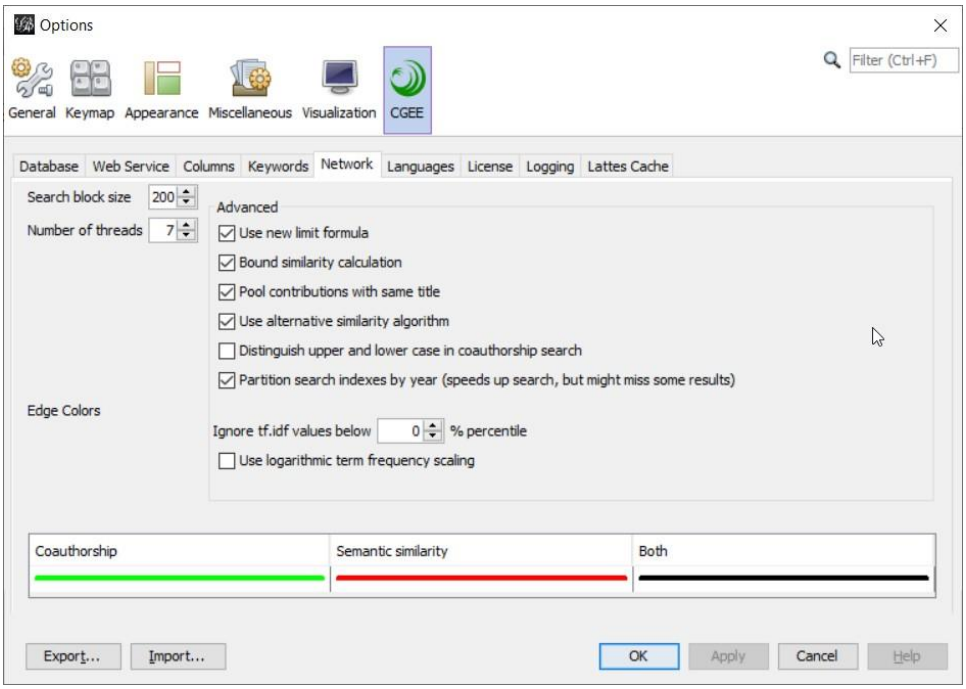


Figura 3.6: Configuração do cálculo e da exibição das redes

Os parâmetros "*Search block size*", "*Number of threads*" e os cinco parâmetros avançados ("*Advanced*") permitem configurar detalhes do processo de pesquisa por coautoria.

"*Search block size*" determina quantas contribuições (artigos, capítulos de livros e trabalhos em eventos) serão agrupados em um bloco de pesquisa, que é alocado a um núcleo de processador do computa- dor. Deve ser considerado que cada bloco gera certo "*overhead*", um processamento adicional. Assim, sugere-se minimizar a quantidade de blocos. Por outro lado, em computadores com vários núcleos, o processamento dos blocos pode ser paralelizado, o que favorece a escolha de uma quantidade maior de blocos. O valor padrão de 200 representa um equilíbrio entre os dois objetivos, mas pode ser alterado pelo usuário.

O parâmetro "*Number of threads*" indica quantos blocos de processamento serão analisados em paralelo. O valor padrão varia de computador para computador. Esse valor é igual à quantidade de processadores (ou núcleos) disponíveis na linguagem Java exceto um, para ainda disponibilizar capacidade de proces- samento para as tarefas de visualização. Caso necessário, esse valor pode ser reduzido para diminuir a carga do computador.

Os parâmetros avançados ("*Advanced*") configuram otimizações dos algoritmos de cálculo de similari- dade. Recomenda-se deixar todos eles ligados para obter o melhor desempenho:

- "*Use new limit formula*": Otimização do cálculo da distância *Levenshtein* máxima a partir da similaridade especifi cada pelo usuário
- "*Bound similarity calculation*": Otimização do critério de conclusão de busca.
- **"*Pool contributions with same title*": Contribuições com o mesmo título são agrupadas. Assim, a** busca por similaridade precisa ser executada apenas uma única vez para todos os títulos iguais.

- "*Use alternative similarity algorithm*": Uso de um algoritmo otimizado de cálculo de similari- dade.
- Geralmente, a pesquisa não distingue entre letras minúsculas e letras maiúsculas e considera pala- vras como "Contribuição" e "CONTRrIBUIÇÃO" como iguais. Marcando a opção "*Distinguish upper and lower case in coauthorship search*", as duas palavras são consideradas diferentes e os resultados dos cálculos de coautoria serão ser diferentes
- Para encontrar contribuições com nome semelhantes, cada tipo de contribuição só é procurada dentro do conjunto de contribuições do mesmo tipo (por exemplo, para achar um artigo com título semelhante, são apenas analisados os títulos dos outros artigos e não dos trabalhos em eventos ou dos capítulos de livros). Esse particionamento reduz significativamente o tempo de busca. Adicionalmente, o usuário pode especificar que as contribuições também devem ter sido publicadas no mesmo ano. Dessa forma, a similaridade de um artigo publicado em 2005 será apenas procurada nos artigos publicados em 2005 (e não nos artigos de todos os anos). Essa opção agiliza significativamente a pesquisa por similaridade, mas pode levar à uma situação onde contribuições inseridas com o ano errado não serão encontradas.
- "*Ignore tf.idf values below x% percentile*": Os termos cujo valor de relevância (tf.idf) são abaixo do percentil confi gurado serão eliminados da busca de similaridade. Se esse valor for diferente de zero, um aviso é exibido no diálogo de busca.
- "*Use logarithmic term frequency scaling*": Nos algoritmos de busca por similaridade semântica, a relevância de um termo é calculado a partir do algoritmo "tf.idf", que considera como um dos seus dois elementos a frequência com que um termo ocorre em um documento. Esta configuração per- mite selecionar se a frequência será utilizada de forma original ou se o logaritmo dessa frequência será usado, que pode ser mais adequado para documentos com tamanhos diferentes. Entretanto, o uso dessa opção deve ser avaliado caso por caso.

3.5.1 Coloração das arestas do grafo

A tabela na parte inferior do diálogo permite a configuração das cores das arestas que aparecem no grafo. As três colunas "*Coauthorship*", "*Contextual similarity*" e "*Both*" mostram as cores em que são exibidas as arestas que possuem apenas coautorias, apenas similaridade contextual ou ambas. Clicando no campo que mostra a linha, uma tela de seleção de cores é exibida:

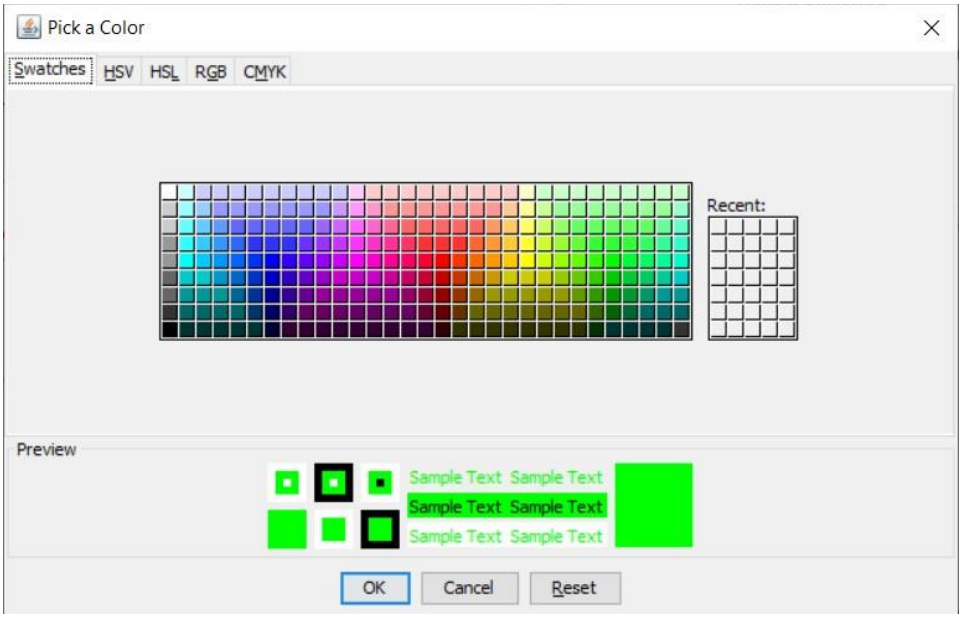


Figura 3.7: Seleção de cores

3.6 Detecção de idiomas

A partir da versão 3.1 do *CGEE Insight Net*, o tratamento de textos em vários idiomas foi reformulado. Para cada documento cuja similaridade será analisada, o *CGEE Insight Net* tenta determinar o idioma do título e do resumo, para poder realizar a análise com os parâmetros corretos do idioma. Essa detecção de idioma pode ser configurado na tela "*Languages*".

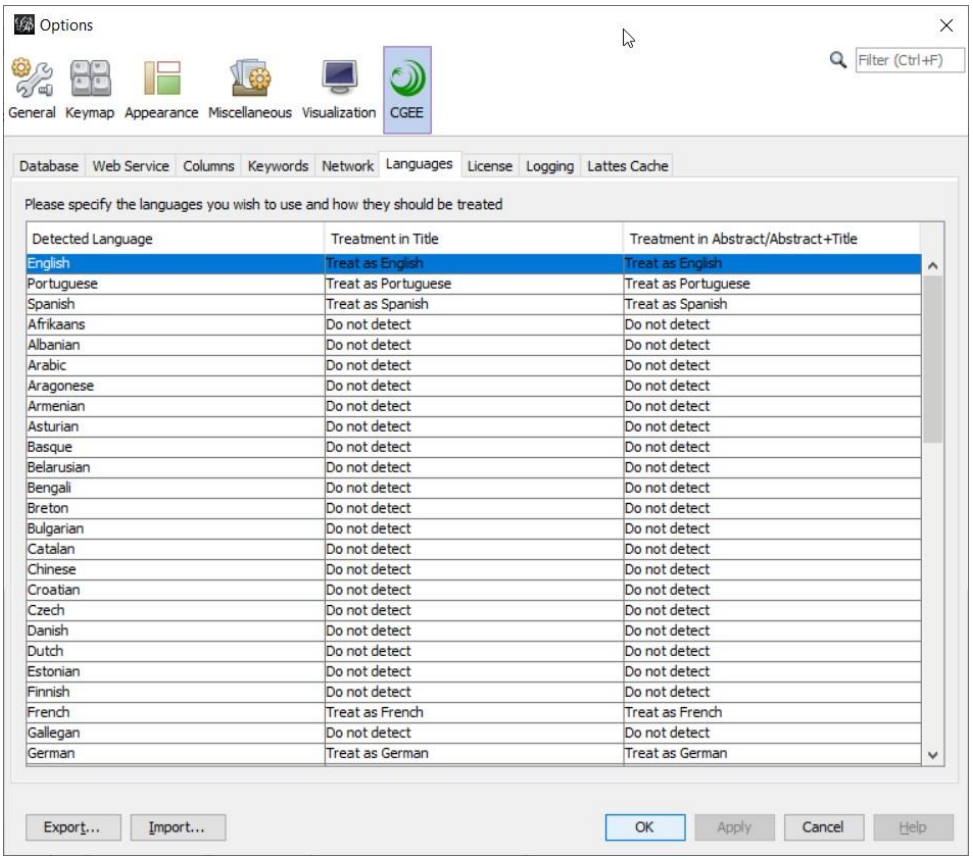


Figura 3.8: Configuração da detecção de idiomas

Deve ser observado que não existem analisadores de texto para todos os idiomas que podem ser detecta- dos. Recomenda-se limitar a quantidade de idiomas detectados para mater a fidelidade dessa detecção. Na configuração básica, são apenas reconhecidos documentos em Inglês, Português, Espanhol, Francês e Alemão. A deteção e o tratamento podem ser diferenciados por título ou por resumo.

3.7 Licenças

O acesso às funcionalidades do *CGEE Insight Net* é restrito por licenças opcionais. Cada módulo possui uma licença específica:

- Currículos Lattes (ver Seção 5)
- Referências bibliográficas BibTeX(ver `bibtex`)
- Referências bibliográficas genéricas (var

Seção 6) Essas licenças podem ser instaladas na

aba "*License*":

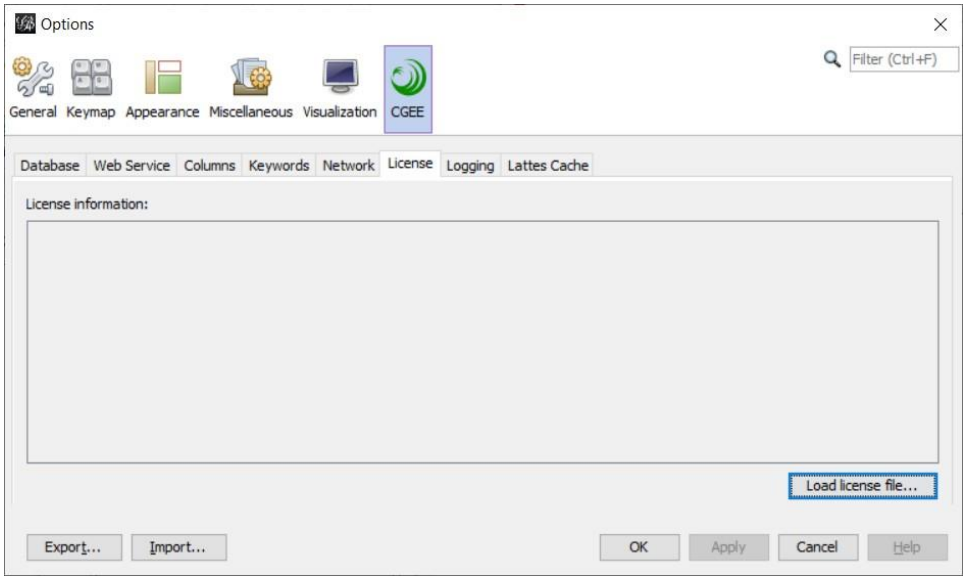


Figura 3.9: Instalação de licenças

As licenças são disponibilizadas em forma de arquivos criptografados que podem ser carregados com o botão "*Load license file*". Depois da validação da licença, a disponibilidade é exibida no diálogo:

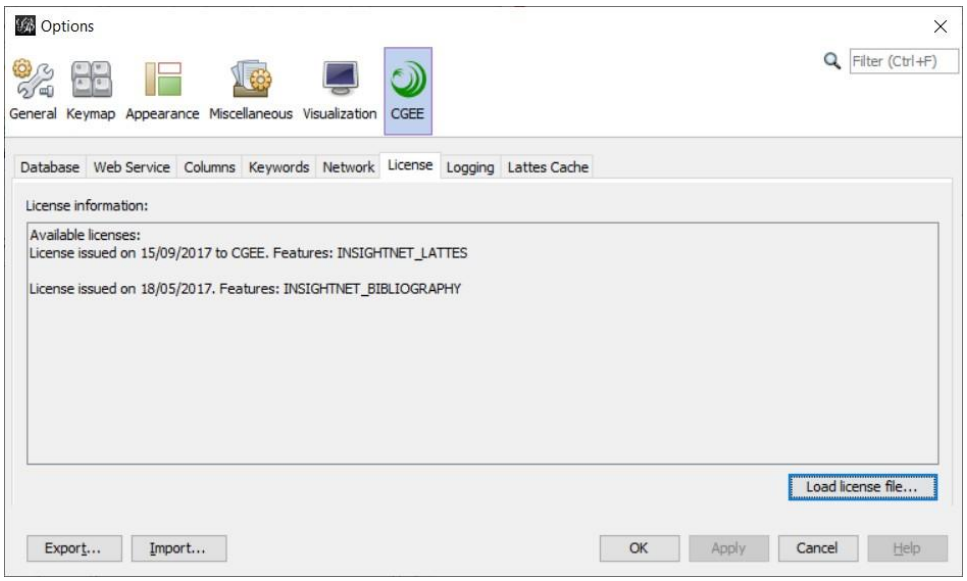


Figura 3.10: Indicação das licenças disponíveis

3.8 Protocolos de execução

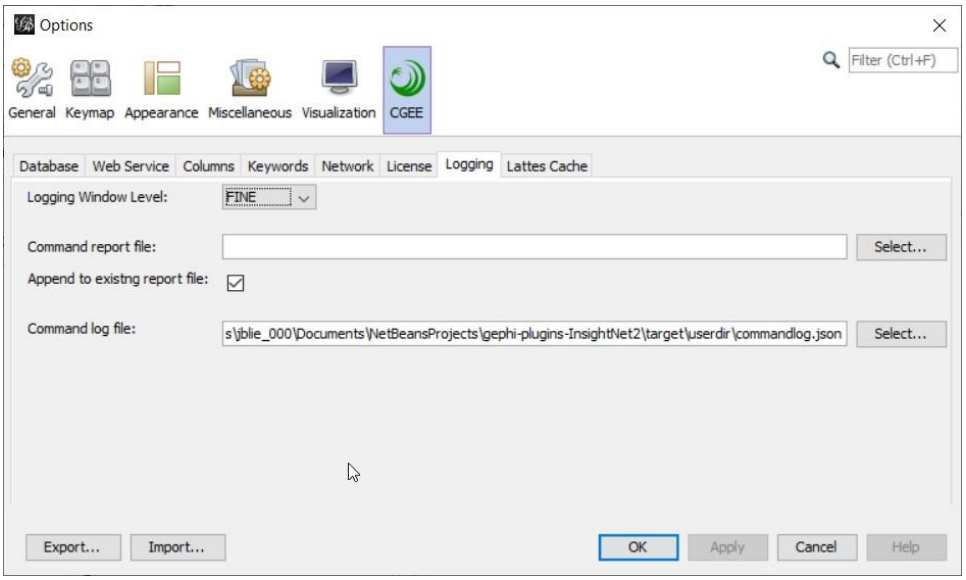


Figura 3.11: Configuração do protocolo de execução

Durante a sua execução, o *CGEE Insight Net* gera informações e avisos que podem ser exibidos na tela e que permitem um acompanhamento e mesmo uma depuração em caso de problemas.

O grau de detalhamento dessas mensagens pode ser configurado com o item "*Logging level*". A configu- ração inicial (*INFO*) gera registros que permitem o acompanhamento da execução no nível de um usuário com pouca experiência. Os níveis *WARNING* e *SEVERE* mostram apenas erros e avisos mais graves e os níveis *FINE*, *FINER* e *FINEST* geram registros de depuração que, geralmente, não são relevantes para os usuários, mas podem ser úteis para a análise de eventuais erros de carga ou de processamento.

Ainda existem dois relatórios de execução, que protocolam as atividades do *plugin*. Enquanto o *Com- mand report file* demonstra as informações da forma legível para o usuário, o *Command log file* é uma protocolo mais adequado para o processamento automático.

3.9 Memória *cache* de Currículos Lattes

O módulo “Currículo Lattes” - caso for habilitado por licença - mantém uma memória local (“*cache*”) para agilizar a carga de currículos recentemente usados. Essa memória é limitada em termos da quanti- dade de currículos, do tamanho total dos currículos e da idade máxima. Esses limtes podem ser confi- gurados nesta aba:

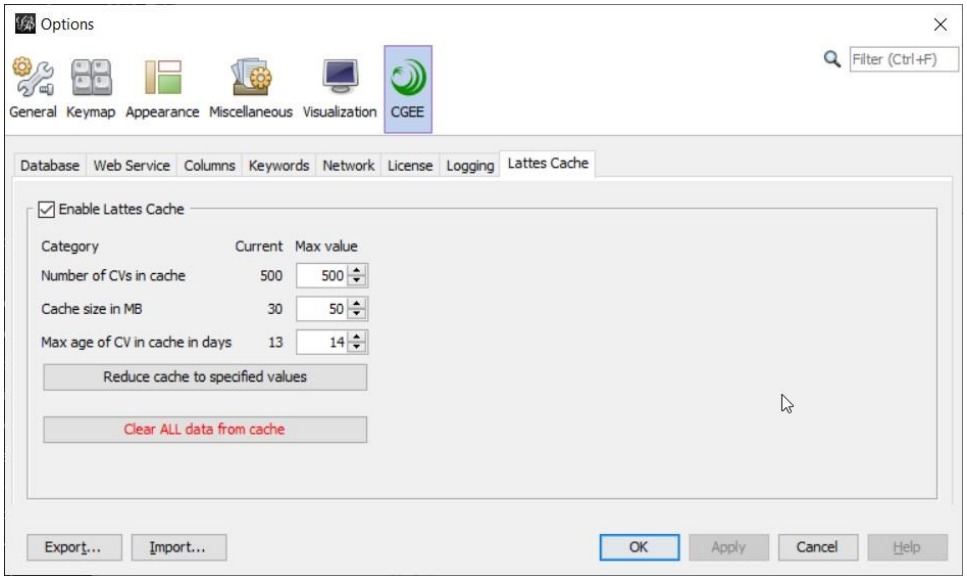


Figura 3.12: Configuração da memória *cache* do módulo Lattes

Nesta aba, a memória *cache* ainda pode ser desabilitada completamente. Ainda podem ser eliminados da memória *cache* todos os currículos ou apenas aqueles que excedem os limites configurados (caso estes valores foram ajustados).

CAPÍTULO 4

Conceitos gerais do uso do *CGEE Insight Net*

4.1 Fluxo de trabalho

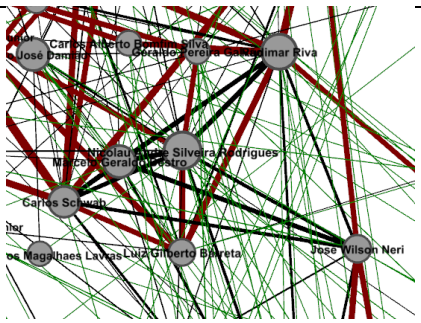
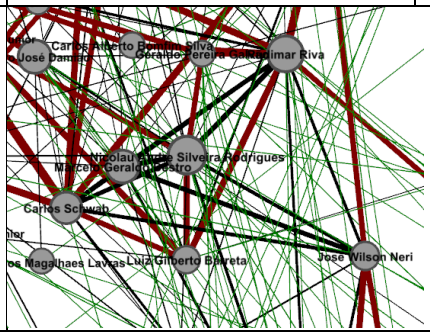
Ao utilizar o *CGEE insight Net*, deve ser considerado que ele trabalha com três repositórios de informa- ções:

- Os dados de entrada, dependendo do módulo de processamento: * Currículos *Lattes* em formato XML * Referências bibliográficas dos sistemas serviços Web of Science® e Scopus® em formato BibTeX * Referências bibliográficas genéricas em formato textual ou planilha Excel®
- O banco de dados contendo todas as informações, incluindo detalhes sobre as contribuições, palavras-chave e graus de similaridade.
- O grafo composto de nós e arestas, ambos com certos atributos.

As informações do grafo podem ser visualizadas e manipuladas diretamente na ferramenta *Gephi*, atra- vés das funções de manipulação de nós e arestas. O acesso ao banco de dados é realizado exclusivamente pelo *CGEE Insight Net*. Essa diferença é importante, pois certas operações feitas no *Gephi* podem não ser registradas no *CGEE Insight Net* e vice-versa.

A tabela em seguida mostra o fluxo típico de informações do processamento no caso dos Currículos *Lattes* que gera uma rede de pesquisadores pelo *CGEE Insight Net*. Para os casos de referência biblio- gráfica, o processo é similar, mas a rede gerada representa contribuições bibliográficas, tais como artigos ou trabalhos em eventos.

Tabela 4.1: Fluxo de informações do *plugin*

Pass	Origem	Operação	Destino
1	<?xml version="1.0" encoding="i <CURRICULO-VITAE SISTEMA-ORIGEM DATA-ATUALIZACAO="10082011" HOR NUMERO-IDENTIFICADOR="854972258 xmlns:lattes="http://www.cnpq.b <DADOS-GERAIS NOME-COMPLETO=" NOME-EM-CITACOES-BIBLIOGRAFIC NACIONALIDADE="B" CPF="147015 UF-NASCIMENTO="MG" CIDADE-NAS Currículo Lattes XML	→ Importação	• Pesquisadores • Artigos • Capítulos Livros • Trabalhos em Eventos • Palavras-chave Banco de dados
2	• Pesquisadores • Artigos • Capítulos Livros • Trabalhos em Eventos • Palavras-chave Banco de dados	→ Seleção de contribuições	• Pesquisadores • Contribuições selecionadas: ○ Artigos ○ Capítulos Livros ○ Trabalhos em Eventos • Palavras-chave Banco de dados
3	• Pesquisadores • Contribuições selecionadas: ○ Artigos ○ Capítulos Livros ○ Trabalhos em Eventos • Palavras-chave Banco de dados	→ Pesquisa de similaridade	• Pesquisadores • Contribuições selecionadas: ○ Artigos ○ Capítulos Livros ○ Trabalhos em Eventos • Palavras-chave • Similaridades • Colaborações Banco de dados
4	• Pesquisadores • Contribuições selecionadas: ○ Artigos ○ Capítulos Livros ○ Trabalhos em Eventos • Palavras-chave • Similaridades • Colaborações Banco de dados	→ Visualização	 Grafo
5		• Pesquisadores • Contribuições selecionadas: ○ Artigos ○ Capítulos Livros ○ Trabalhos em Eventos • Palavras-chave • Similaridades • Colaborações Banco de dados	Análise do grafo e pesquisas de palavra-chave

Esta sequência demonstra que o grafo é gerado apenas no último passo de visualização. É relevante mencionar que manipulações no grafo, que são operações do *Gephi*, não se

refletem dentro do banco de dados. Por outro lado, podem ser geradas várias visualizações do mesmo banco de dados, permitindo

análises visuais diferentes a partir do mesmo banco de dados. A separação do grafo do banco de dados também permite o compartilhamento de dados no nível de rede (nós e arestas) sem divulgar dados potencialmente sigilosos que constam nos currículos importados.

As seções a seguir detalham os passos descritos.

CAPÍTULO 5

Uso do *CGEE Insight Net* para analisar Currículos Lattes

O CGEE Insight Net permite a criação de redes de pesquisadores por co-autorias e similaridade semân- tica das publicações.

Para habilitar essa funcionalidade do *CGEE Insight Net*, a licença `INSIGHTNET_LATTES` deve ser instalada, conforme descrito na Seção 3.7. Essa licença é exibida assim no diálogo *Tools* > *Options* > *CGEE* > *License*:

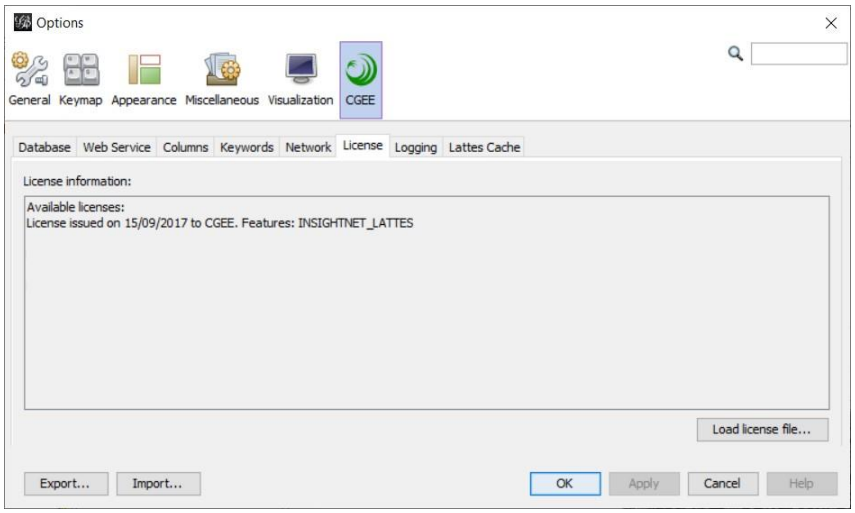


Figura 5.1: Licença requerida para o módulo de redes de

Currículos Lattes Caso a licença esteja habilitada, aparece o sub-menu *"CGEE Insight Net Lattes"*:

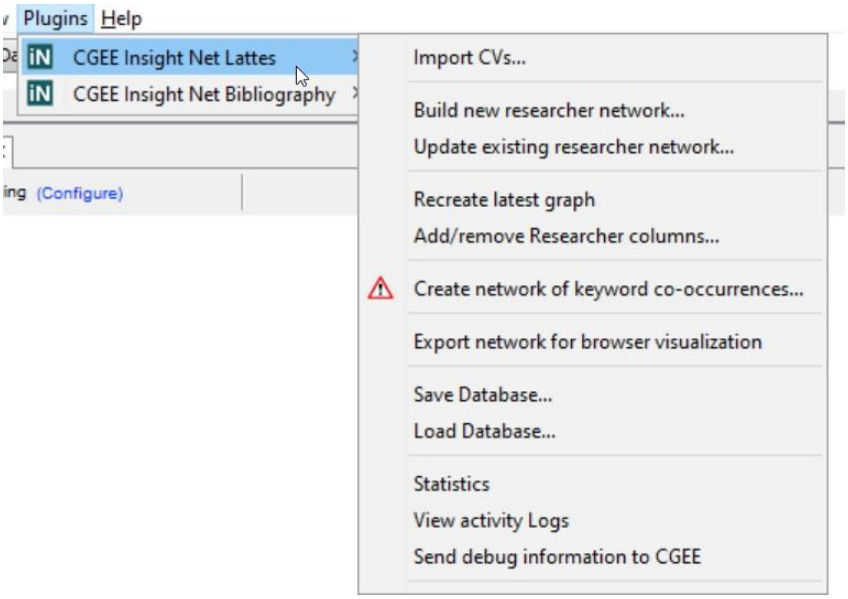


Figura 5.2: Sub-menu CGEE Insight Net Lattes

5.1 Importação dos Currículos Lattes

Para processar as informações dos Currículos Lattes em formato XML, estes devem ser importados no banco de dados a partir da função *Plugins* > *CGEE insight Net Lattes* > *Import*, que exibe o seguinte diálogo, cujo formato depende da aba selecionada na parte superior:

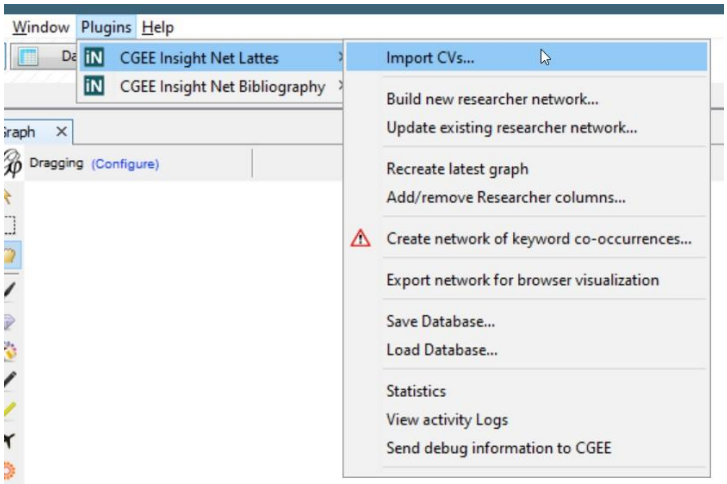


Figura 5.3: Menu de importação de Currículos

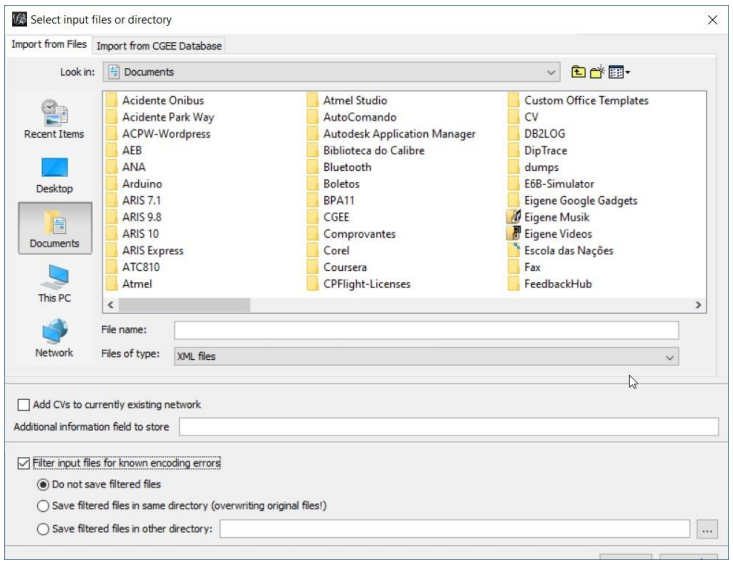


Figura 5.4: Diálog de importação de Currículos em Arquivos

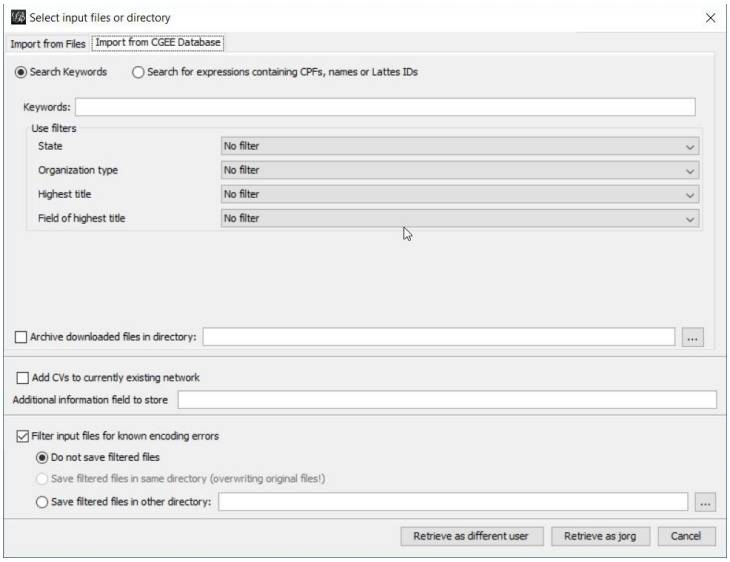


Figura 5.5: Diálogo de importação de Currículos do Banco de Dados

Esse diálogo permite a importação de currículos Lattes em arquivos XML ou por conexão direta com o banco de dados do CGEE.

5.1.1 Importação de arquivos XML

Selecionando a aba "*Import from files*", o usuário pode importar currículos gravados no formato XML no computador local, em algum diretório compartilhado em rede ou mesmo um dispositivo móvel de armazenamento. O escopo da importação depende da seleção dos arquivos na lista apresentada:

- Clicando em um arquivo XML, este será importado;
- Vários arquivos XML podem ser selecionados com "*Shift-Clique*" ou "*Ctrl-Clique*, de acordo com os padrões de uso do sistema operacional;
- O usuário também pode selecionar um ou mais diretórios. Nesse caso, todos os arquivos XML nesses(s) diretório(s) serão importados.

Os arquivos importados devem seguir o padrão XML dos Currículos Lattes do CNPq[1].

5.1.2 Acessando o banco de dados do CGEE

Selecionando a aba "*Import from CGEE database*", o software pode acessar diretamente o banco de da- dos de currículos do CGEE. Neste caso, o diálogo oferece duas funcionalidades para recuperar currículos Lattes. Essas funcionalidades serão descritas em seguida.

[1] Verificar em http://lattes.cnpq.br/web/plataforma-lattes/extracao-de-dados

Recuperação por palavras-chave

A opção "*Search Keywords*" permite que o usuário especifique palavras-chave que serão aplicadas na pesquisa de especialistas por competência, seguindo padrões de uso do Portal da Inovação[2]. Adicional- mente, os currículos obtidos podem ser filtrados por "Unidade da Federação", "Tipo de organização", "Maior titulação" e "Área da maior titulação":

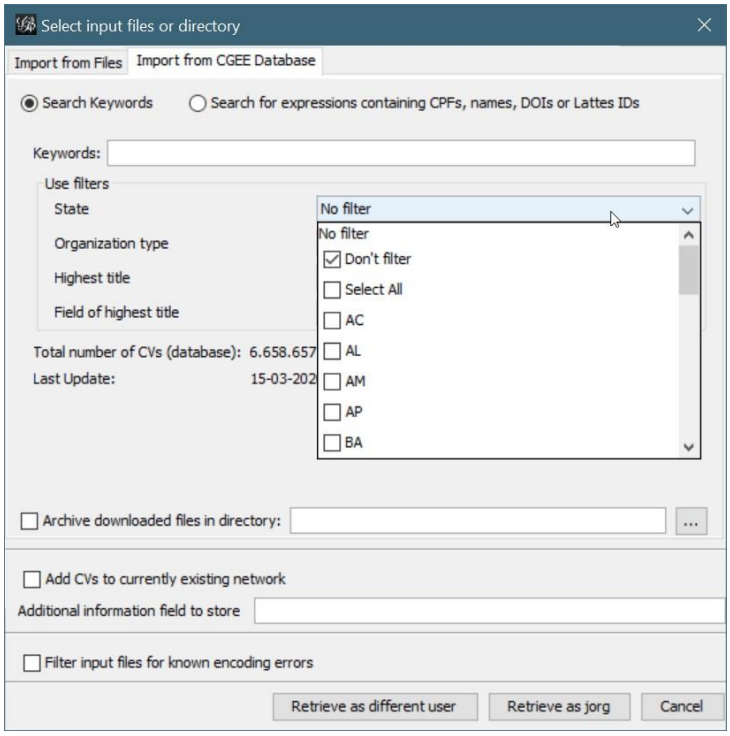


Figura 5.6: Importação de dados do Banco de

dados do CGEE Destaca-se a diferença entre as primeiras duas

opções de cada filtro:

- "*Don't filter*" não aplica nenhum filtro nos currículos
- "*Select all*" elimina aqueles currículos cujo critério não consta na lista de valores válidos

Como exemplo, é possível citar o pesquisador estrangeiro que não preencheu o campo "UF" no seu currículo. Se o usuário escolher "*Select all*", este currículo não fará parte da importação, pois "*Select all*" considera apenas currículos que possuem um dos valores definidos na lista de estados. Para incluir o pesquisador estrangeiro, o usuário teria que selecionar "*Don't filter*".

Durante a digitação da palavra-chave, uma busca prévia é iniciada no servidor e a

quantidade de cur- ríc ulos que atendem aos critérios selecionados é exibida no campo "*No. of found CVs*". Para isso, é necessário que o usuário tenha digitado, no mínimo, três letras no campo "*Keywords*", seguido por um intervalo sem digitação de, no mínimo, dois segundos.

As palavras-chave digitadas são automaticamente copiadas para o campo "*Additional information field to store*" e serão exibidas no campo "info" do Laboratório de dados do Gephi.

[2] Verificar em http://www.portalinovacao.mcti.gov.br/pi/#/pi

Pesquisa por filtro

A opção "*Search for expressions containing CPFs, names or Lattes IDs*" permite a especificação de filtros usados pelo software "*WebExtractor*" do CGEE:

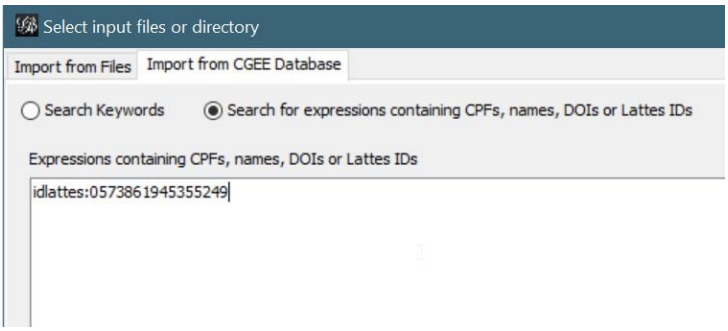


Figura 5.7: Recuperação de currículos por

expressão de busca A expressão do filtro deve usar o seguinte

formato:

```
<critério de busca>:<item1>,<item2>,...
```

O critério de busca determina o campo de informação usado para realizar as buscas e pode ser um dos seguintes valores:

- CPF: realiza uma busca pelos CPFs dos pesquisadores
- Nome: realiza uma busca pelos nomes dos pesquisadores, sem acentos e caracteres especiais
- Idlattes: realiza uma busca pelo número identificador do currículo na

base Lattes Seguem alguns exemplos de expressões de filtros:

- Extração dos currículos dos pesquisadores que possuem os CPFs 123.456.789-12 ou 987.654.321- 00: `cpf:12345678912,98765432100`
- Extração do currículo do pesquisador "Pesquisador 1": `nome: Pesquisador 1`
- Extração dos currículos dos pesquisadores "Pesquisador 1", "Pesquisador 2" e "Pesquisador 3":
  `nome:Pesquisador 1,Pesquisador 2,Pesquisador 3`
- Extração dos currículos Lattes com os identificadores "0000000000000000" e "1111222233334444": `idlattes:0000000000000000,1111222233334444`

Arquivamento dos dados originais

Os currículos Lattes recuperados podem ser arquivados, junto com um texto descritivo da

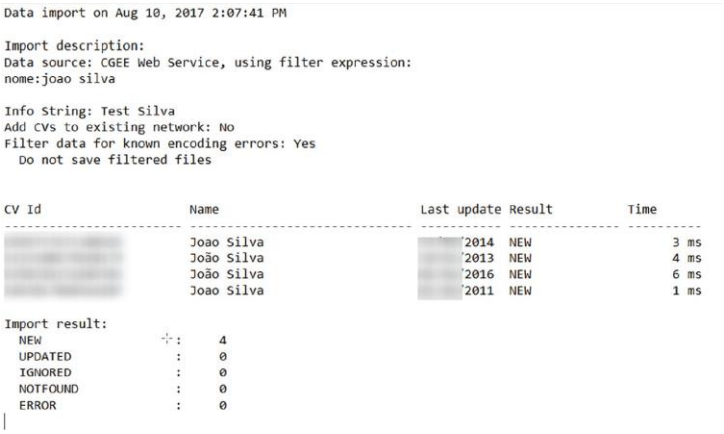
operação. Essa funcionalidade, importante para evitar que atualizações do Lattes inviabilizem a reprodução e validação de uma análise realizada, é ativada com a opção "*Archive downloaded files in directory*":

Figura 5.8: Arquivamento dos dados originais

Se essa opção estiver ligada e um diretório válido for especificado, cada importação gerará um arquivo

`Import _ <data> _ <hora>.zip` que contém

- Todos os currículos baixados, bem como
- Um arquivo `SUMMARY.TXT`, que descreve os insumos e o resultado da operação



```
Data import on Aug 10, 2017 2:07:41 PM

Import description:
Data source: CGEE Web Service, using filter expression:
nome:joao silva

Info String: Test Silva
Add CVs to existing network: No
Filter data for known encoding errors: Yes
  Do not save filtered files

CV Id                  Name                      Last update Result         Time
---------------------- ------------------------- ----------- -------------- ---------
[illegible]            Joao Silva                [illegible]2014 NEW                3 ms
[illegible]            João Silva                [illegible]2013 NEW                4 ms
[illegible]            João Silva                [illegible]2016 NEW                6 ms
[illegible]            Joao Silva                [illegible]2011 NEW                1 ms

Import result:
  NEW                  :     4
  UPDATED              :     0
  IGNORED              :     0
  NOTFOUND             :     0
  ERROR                :     0
```

Figura 5.9: Arquivo SUMMARY.TXT, descrevendo os insumos e o resultado da importação

5.1.3 Opções comuns

As opções descritas em seguida permitem controlar o processo da importação, independentemente da fonte de dados.

Apagar ou manter os dados do banco antes da importação

A opção "*Add CVs to currently existing network*" está disponível se, na hora da importação, já houver um banco de dados com currículs Lattes e diferencia entre uma importação inicial e uma importação incremental (que não elimina dados anteriores).

Figura 5.10: Opção de importação inicial ou incremental

Se essa opção for selecionada, os currículos importados serão acrescentados às informações já existentes na base. Se um currículo importado já existe na base e a versão importada é mais recente do que a versão na base, o currículo na base é substituído pela versão importada.

Se a opção não for selecionada, todos os dados que já existem no banco de dados serão apagados antes da importação. Desta forma, os dados importados substituem os dados existentes.

Campo adicional de informação

Cada pesquisador importado é representado como um nó no grafo criado. Esses nós possuem atributos, tais como o número do Currículo Lattes (atributo “id”), o nome do pesquisador (atributo “*label*”) e ou- tros. O atributo “*info*” dos nós é preenchido com o valor especificado no campo “*Additional information field to store*” durante a importação.

Figura 5.11: Opção do campo adicional de informação

Essa funcionalidade permite que durante a importação incremental de vários Currículos Lattes em vários passos os pesquisadores sejam categorizados em grupos com identificadores distintos. Se o mesmo cur- rículo é importado várias vezes com valores diferentes no campo “info”, esses valores serão adicionados e separados com o caractere “/”.

Limpeza dos dados

Para permitir o processamento de arquivos contendo alguns tipos de erros identificados na base de cur- rículos, foi desenvolvida uma correção automática, da seguinte forma:

- Caracteres CTRL-Z são substituídos por símbolos de interrogação (“?”).
- A codificação dos arquivos é determinada automaticamente e, caso não confira com a codificação declarada na linha inicial, a declaração é corrigida.

Figura 5.12: Limpeza dos dados e gravação dos arquivos corrigidos

Se a limpeza dos dados for habilitada com a opção “*Filter input files for known encoding errors*”, os currículos corrigidos podem opcionalmente ser gravados na mesma pasta (sobrescrevendo os arquivos originais) ou em outra pasta.

5.1.4 Processo de importação

Durante a importação, o CGEE Insight Net mostra uma barra de progresso e informa sobre o andamento da importação.

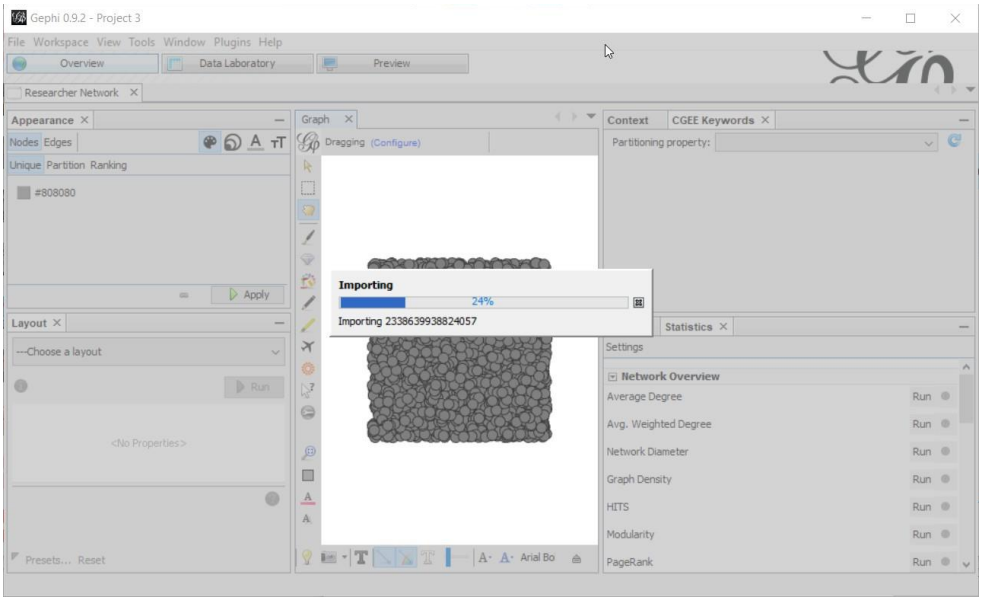


Figura 5.13: Importação dos currículos Lattes

No final da importação, a quantidade de pesquisadores importados, atualizados, ignorados e não impor- tados por erros nos dados é exibida:

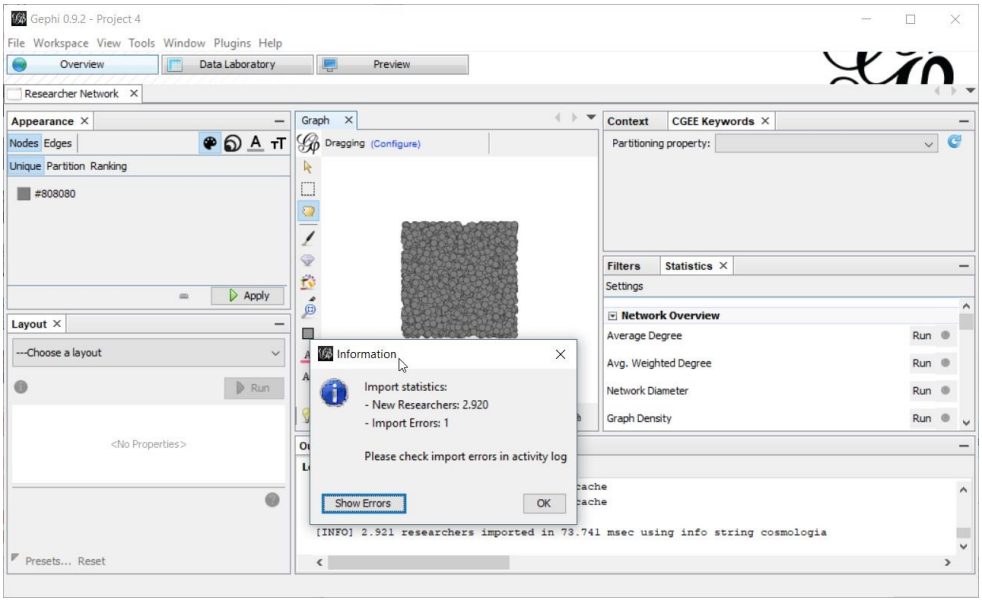


Figura 5.14: Resultado da importação dos currículos Lattes

Recomenda-se verificar essa quantidade de pesquisadores com a quantidade esperada para identificar possíveis divergências.

Caso forem identificados erros na importação, esses podem ser exibidos e detalhados. Adicionalmente, os currículos correspondentes podem ser baixados diretamente do site do CNPq em formato XML e importados manualmente (ver seção Seção 5.1.1).

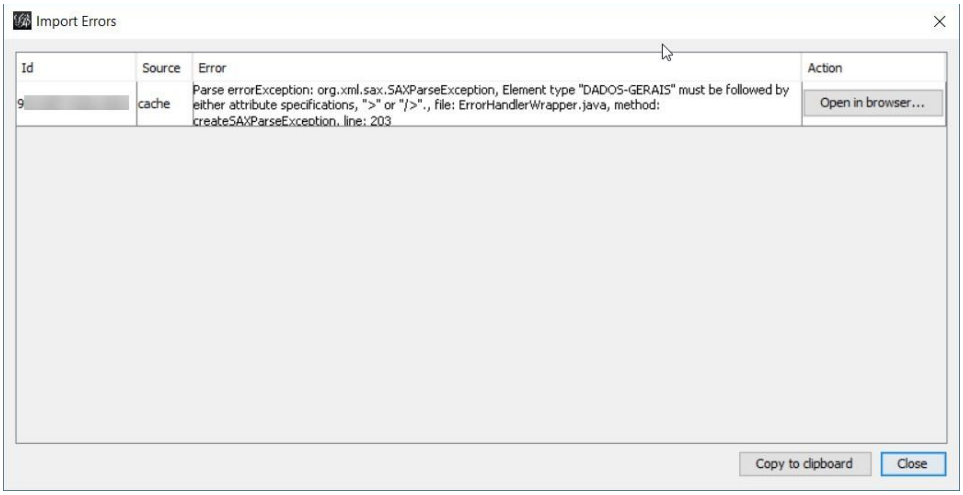


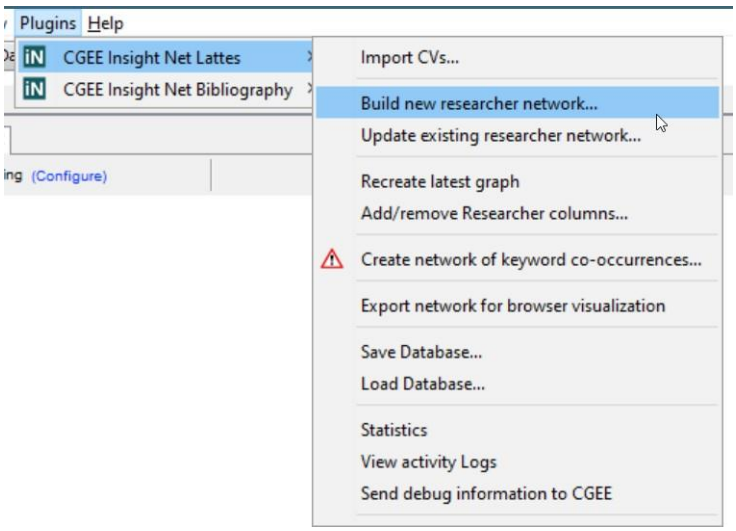
Figura 5.15: Diálogo de erros de importação

Adicionalmente, o protocolo de execução (ver seção *Protocolos de execução*) registra informações sobre o andamento da importação, de acordo com o grau de detalhe especificado na tela de configuração (ver seção :*options-log*).

O relatório de execução (ver seção *Protocolos de execução*) reune todas as informações detalhadas da importação, no mesmo formato do arquivo `SUMMARY.TXT` (veja *Arquivamento dos dados originais*).

5.2 Formação da rede

Depois da importação dos currículos na base de dados, a rede é formada a partir das pesquisas por coautoria e por similaridade contextual. Os passos 2-4 da Tabela Tabela 4.1 são realizados em uma única operação, transformando o conteúdo do banco de dados em um grafo. Para formar a rede, o usuário deve clicar em *Plugins* > *CGEE Insight Net Lattes* > *Build new researcher network* e preencher ou confirmar os dados do diálogo que é exibido:



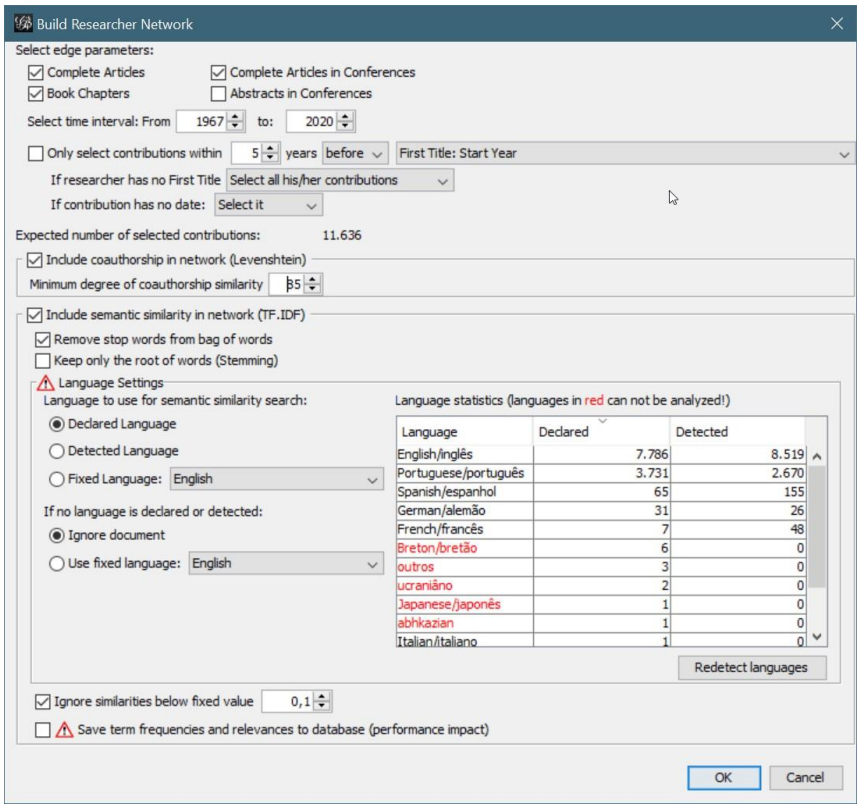


Figura 5.16: Menu e diálogo da formação da rede de

currículos Lattes As opções do diálogo serão explicadas em seguida.

5.2.1 Escopo da rede formada

Na parte superior do diálogo o usuário especifica quais tipos de contribuições farão parte do escopo da formação da rede:

- Artigos científicos completos (desconsiderando artigos de resumo em eventos)
- Capítulos em livros
- Trabalhos em eventos (Artigos completos ou apenas Resumos)
- As contribuições selecionadas podem ainda ser limitadas por período de publicação – o diálogo mostra o ano mínimo e o ano máximo de todas as contribuições importadas.
- Outra possível limitação do escopo temporal é em relação às titulações dos pesquisadores. Se a caixa "*Only select contributions within _____ years*" for selecionada, apenas as contribuições dentro da faixa temporal selecionada farão parte da rede construída. Nesse caso, é necessário definir o tratamento das contribuições dos pesquisadores que não obtiveram a titulação selecionada e também das contribuições que não possuem data.

A visualização dos detalhes das contribuições Lattes (ver Seção 5.3) permite a inclusão e exclusão manual de contribuições bibliográficas no escopo de cálculo da rede. Caso for realizada alguma mudança de seleções desta forma, a seguinte informação é exibida pelo *plugin*:

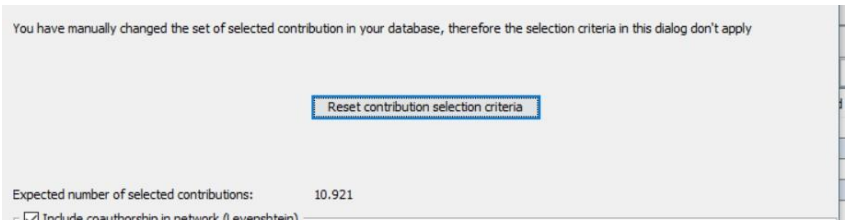


Figura 5.17: Informação sobre a alteração manual do escopo de cálculo da rede

Para desfazer as alterações manuais do escopo de rede, o usuário pode clicar no botão `*Reset contri- bution selection criteria*`. Com esta ação, o escopo de cálculo da rede volta aos critérios algoritmicos mencionados em cima.

Um fato relevante é que a rede dos pesquisadores será montada **apenas** pelas contribuições aqui seleci- onadas.

A quantidade de contribuições selecionadas é calculada quando o diálogo for aberto e cada vez quando uma das opções mencionadas é alterada. Durante o tempo desse cálculo, as opções de seleção permane- cem desabilitadas:

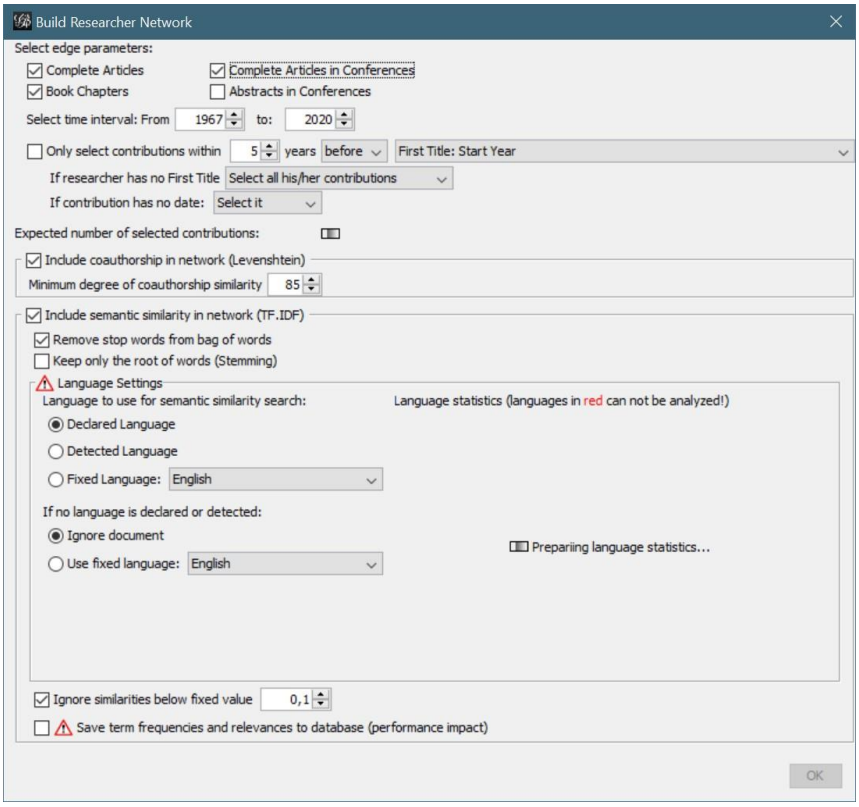


Figura 5.18: Recálculo da quantidade de contribuições selecionadas

5.2.2 Opções da pesquisa por coautoria

Na parte do meio do diálogo, existem algumas opções que permitem controlar o processo de pesquisa por coautoria:

- A caixa "*Include coauthorship in network (Levenshtein)*" permite determinar se a pesquisa por coautoria é realizada e habilita os outros campos desta seção. Se essa caixa não for selecionada, a rede formada não terá arestas de coautorias.
- A similaridade mínima a partir da qual os títulos de duas contribuições são considerados iguais também pode ser alterada pelo usuário. Valores entre 85% e 90% se mostraram adequados para minimizar a incidência de falsos positivos e falsos negativos no que se refere ao número de coau- torias.

5.2.3 Opções da pesquisa por similaridade semântica

A parte inferior do diálogo permite a seleção das opções de pesquisa por similaridade semântica:

- A caixa "Include semantic similarity in network (TF.IDF)" determina se a pesquisa por simila- ridade semântica (também conhecida como "Similaridade contextual") é realizada e habilita os outros campos dessa seção. Se essa caixa não for selecionada, a rede formada não terá arestas de similaridade semântica.
- O usuário pode selecionar se os pré-processamentos dos termos "*Stop words*" e "*Stemming*" serão realizados ou não.

- "*Stop Words*" são as palavras mais frequentes de cada idioma, que não agregam informação aos termos identificados e serão eliminados da pesquisa. Os *stop words* são implementados apenas para os títulos em Inglês e Português.
- O "*Stemming*" reduz, em um algoritmo específico por idioma, cada palavra a uma raiz que desconsidera flexões gramaticais. Nesse momento, apenas os idiomas Português e Inglês são tratados pelo stemming. Títulos em outros idiomas permanecem na forma original.
- Se qualquer uma dessas opções for selecionada, o idioma do texto se torna relevante. Neste caso, aparece no diálogo a estatística de idiomas declarados e detectados nas contribuições.

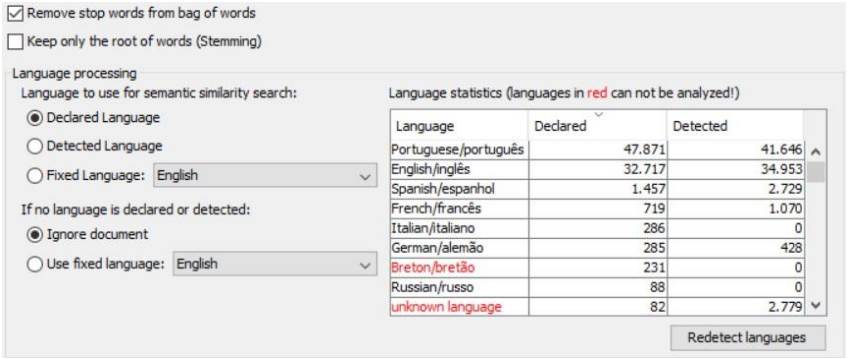


Figura 5.19: Estatística de idiomas detectados e declarados

- O botão "*Redetect languages*" permite realizar uma nova detecção de idiomas com parâme- tros diferentes daqueles configurados na tela "*Languages*" da configuração do plugin (ver Seção 3.6):

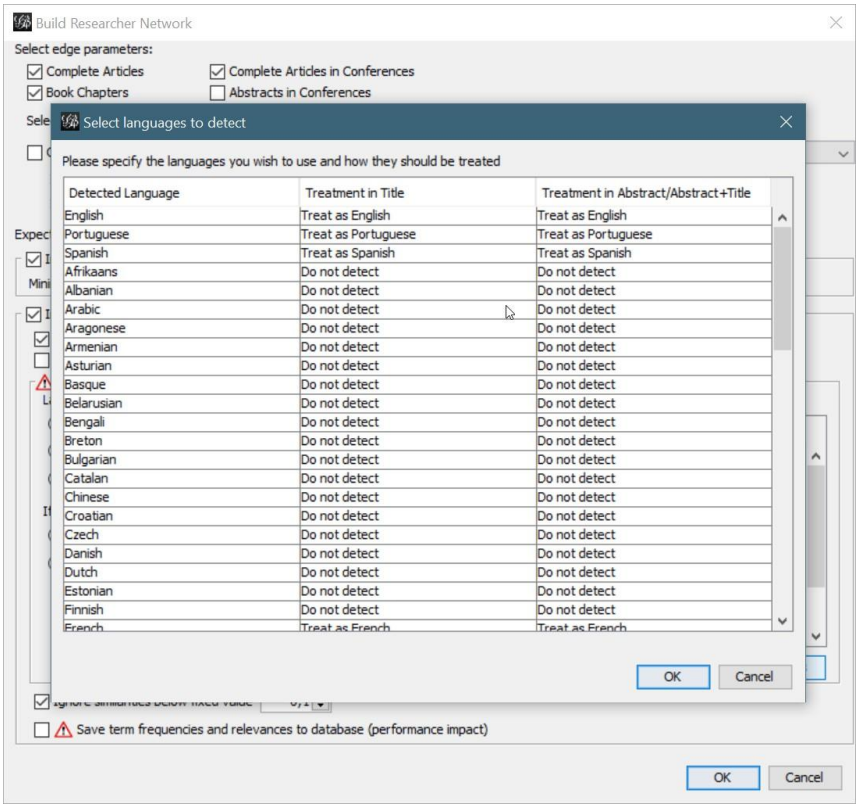


Figura 5.20: Nova detecção de idiomas

- O cálculo da similaridade semântica gera arestas entre praticamente qualquer par de pesquisado- res. A grande maioria deles com baixos valores de similaridade que não agregam informações relevantes ao conteúdo do gráfico. Por esse motivo, existem três métodos para reduzir a quantida- des de arestas na rede:
  - "*Ignore similarities below fixed value*": valores abaixo de um limite especificado podem ser desconsiderados, produzindo o valor final zero como similaridade contextual.
  - "*Sparsify network automatically*": Um algoritmo automático [7] é utilizado para reduzir a quantidade de arestas na rede. Observe-se que testes realizados com esse algoritmo não levaram a resultados conclusivos quanto à sua eficácia.
  - Para o cálculo de similaridade podem ser considerados apenas os termos mais relevantes das contribuições. Essa configuração é realizda no diálogo de opções do *CGEE Insight Net* (ver seção *Configuração do CGEE Insight Net*), advertindo-se que, na grande maioria dos casos, essa opção só deva ser empregada por usuários experientes. Se um percentil de relevância dos termos for definido, uma mensagem correspondente é exibida:

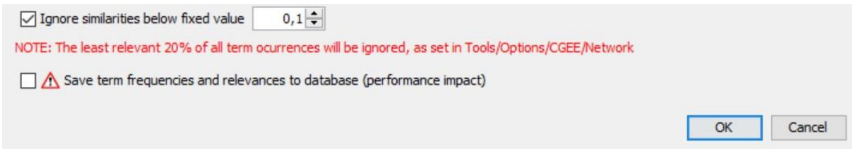


Figura 5.21: Aviso sobre configuração de limite inferior de relevância

Clicando em "OK", o *CGEE Insight Net* inicia a sequência de processamento: No primeiro passo, as contribuições dentro do escopo especificado são identificadas, selecionadas e pré-processadas. O processo pode ser interrompido clicando no símbolo do indicador de progresso:

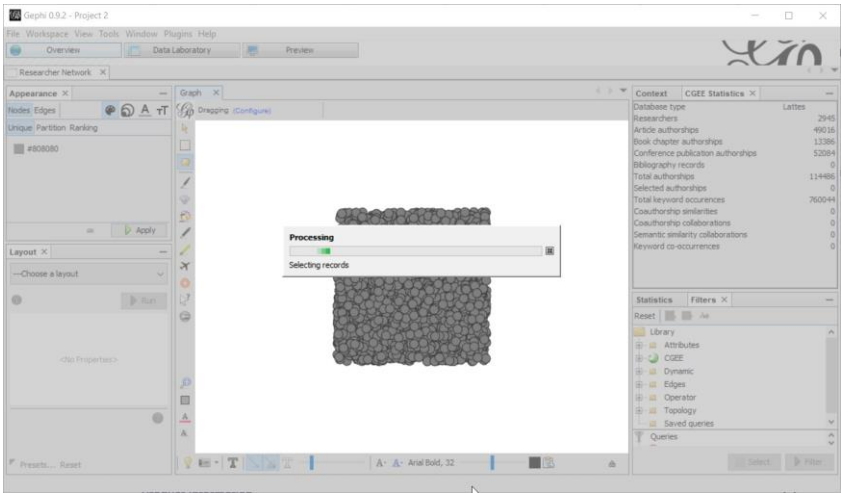


Figura 5.22: Seleção das contribuições e pré-processamento

Depois desta fase, o *CGEE Insight Net* inicia a formação da rede (passo 3 da Tabela 4.1). Esse passo pode levar um tempo considerável, dependendo da quantidade de contribuições

selecionadas, da capacidade do computador de paralelizar a pesquisa (quantidade de processadores e núcleos), do tipo do banco de dados, da velocidade de conexão e dos parâmetros especificados pelo usuário. O *CGEE Insight Net* mostra uma barra de progresso que indica o porcentual dos dados já processados. Novamente, o processo pode ser interrompido clicando no símbolo desta barra:

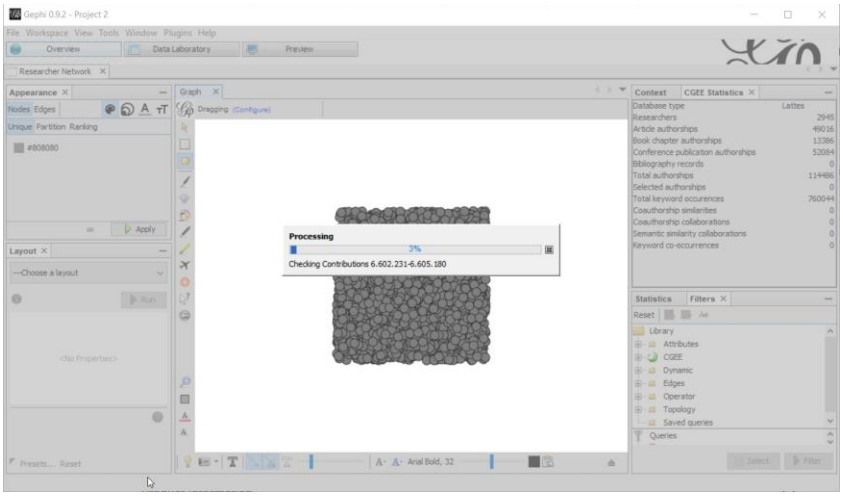


Figura 5.23: Processamento da pesquisa por similaridade

Depois da conclusão deste passo, os dados são pós-processados e a rede de colaboração é montada visualmente na tela:

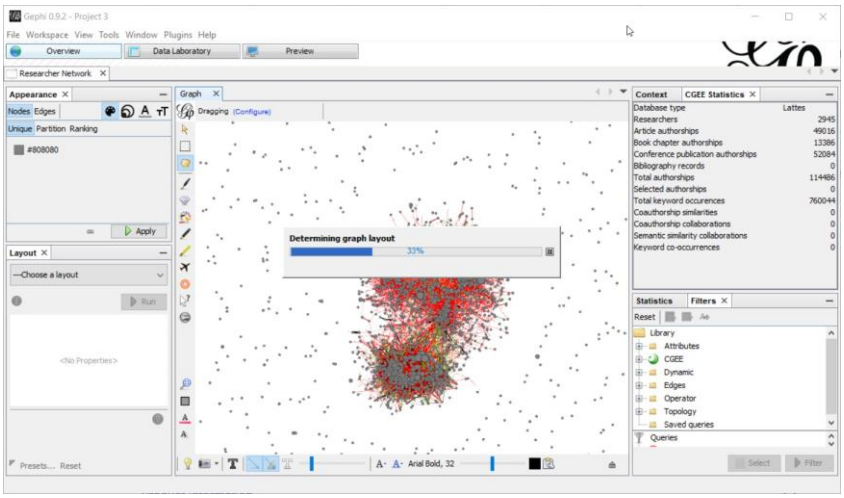


Figura 5.24: Pós-processamento e montagem da rede na tela

Após a conclusão desta etapa, a tela é liberada pelo *CGEE Insight Net* e o usuário pode analisar a rede com as ferramentas disponíveis do Gephi, tais como análise de *clusters*, particionamentos ou estatísticas da rede, usando as ferramentas disponibilizadas pelo Gephi.

5.2.4 Atualização da pesquisa

Para atualizar os cálculos de uma rede formada, o usuário pode selecionar o item *"Update existing researcher network. . . "*:

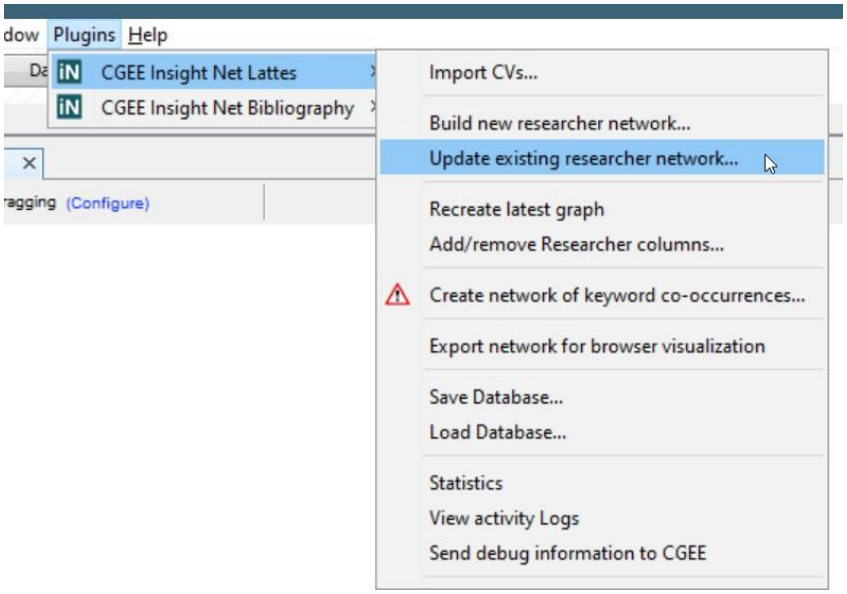


Figura 5.25: Atualização da rede já calculada

Diferentemente da opção *"Build new researcher network."*, que elimina qualquer rede pré-existente, o item *"Update existing researcher network."* mantém os dados dos cálculos que não são selecionados pelo usuário. Dessa forma, é possível atualizar apenas as arestas de similaridade contextual, sem recalcular toda a rede de coautorias.

5.3 Visualização de atributos dos pesquisadores

Cada pesquisador possui uma grande quantidade de atributos que são obtidos através do seu Currículo Lattes:

- Nome completo;
- Nome em citações;
- Instituição;
- Estado da Instituição;
- Quantidade de artigos completos, capítulos de livros e publicações completas e resumos em even- tos;
- Ano da última atualização do currículo;
- Campo adicional de informação, definido pelo usuário durante a importação dos dados;
- Quantidade total, bem como de cada tipo de contribuição bibliográfica;
- Data de nascimento;
- Local de nascimento;
- Dados da primeira, da última e da mais alta titulação que constam no currículo: ano de início e de fim, ano da titulação, tipo, instituição e assunto da titulação.
- Os mesmos dados são registrados para a primeira e última titulação de cada tipo de titulação: - Ensino fundamental; - Ensino médio; - Curso técnico profissionalizante; - Graduação; - Especi- alização; - Residência Médica; - Mestrado profissionalizante; - Mestrado; - Doutorado; e - Livre Docência.

Os seguintes atributos estão disponíveis apenas por pessoas com autorização explícita para processar dados pessoais:

- CPF;
- Sexo.

Os seguintes atributos são definidos para cada pesquisador durante o cálculo de uma rede de co-autorias ou de similaridade semântica:

- Quantidade total, bem como de cada tipo de contribuição bibliográfica **selecionada para o cálculo da rede**;
- Quantidade média de palavras chave por contribuição bibliográfica **selecionada para o cálculo da rede**;
- Porcentagem das contribuições bibliográficas selecionadas para o cálculo de rede que possuem, no mínimo uma palavra-chave.

O cálculo de uma rede de similaridade semântica ainda fornece os seguintes atributos:

- Quantidade total de *termos* em todas as contribuições bibliográficas selecionadas do pesquisador;
- Quantidade de **termos diferentes** em todas as contribuições bibliográficas selecionadas do pes- quisador.

Um *termo* no sentido do parágrafo anterior corresponde, básicamente a uma plavra. Entretanto *stop- words* (ver Seção 5.2.3) não fazem parte dessa contagem e o *stemming* pode reduzir a quantidade de termos diferentes.

A relevância dessas informações depende do projeto específico do usuário, que deve escolher o subcon- junto que melhor atende aos seus requisitos.

Os dados relevantes podem ser selecionados com a função *"Add/remove Researcher columns. . . "*:

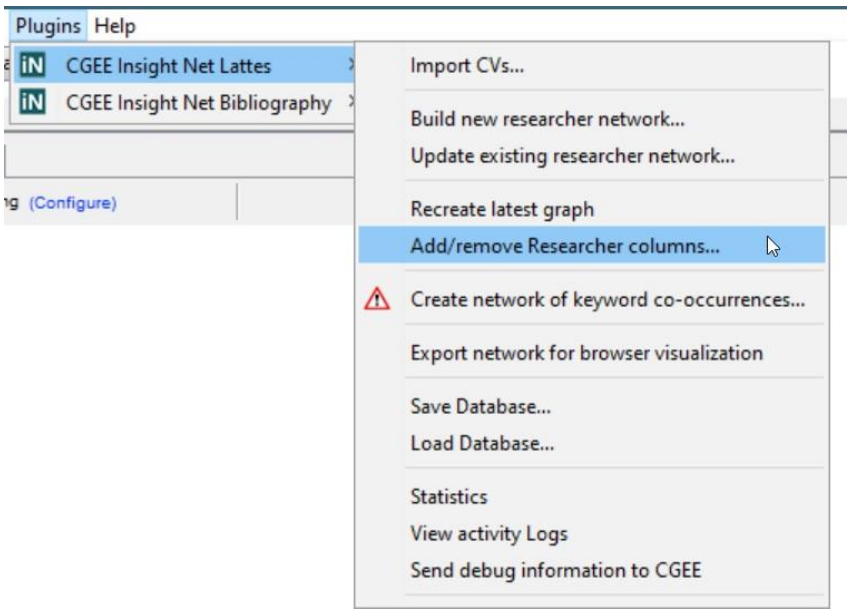


Figura 5.26: Funcionalidade para selecionar dados relevantes dos

pesquisadores Na seleção dessa funcionalidade, o seguinte diálogo é exibido:

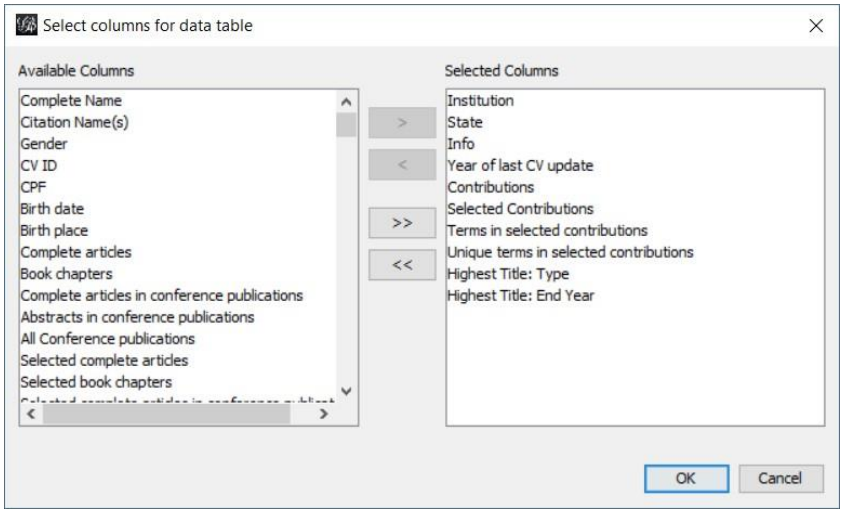
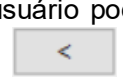
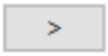


Figura 5.27: Seleção dos atributos dos pesquisadores

O diálogo corresponde à tela de opções descrita na Seção 3.3. A lista à direita mostra os atributos atualmente exibidos e a do lado esquerdo contém os atributos disponíveis (não exibidos). O usuário pode selecionar atributos nas listas com clique e *Shift/Ctrl-clique* e usar os botões 

, 

,

<< e >> para movê-los entre as duas listas, conforme descrito na Seção 3.3.

Destaca-se que o Gephi possui duas funcionalidades de visualização dos atributos. A primeira é na exibição do grafo da rede. Selecionando a ferramenta e clicando em um dos nós do grafo, todos os atributos desse nó serão exibidos, desde que eles tenham sido selecionados com a função *"Add/ Remove Researcher column"* ou no diálogo *"Tools/Options/CGEE/Columns"* do Insight Net Browser:

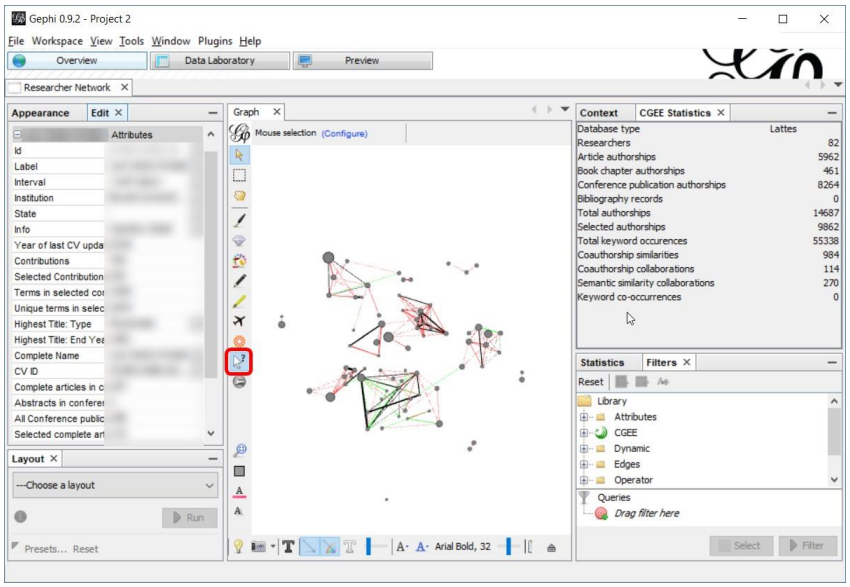


Figura 5.28: Exibição no grafo de todos os atributos habilitados do pesquisador

Na alternativa de visualização do laboratório de dados (*"Data Laboratory"*), o Gephi limita a quantidade de colunas exibidas em 20. Se os nós da rede apresentarem mais atributos, os 20 mais relevantes devem ser selecionados com um clique no símbolo :

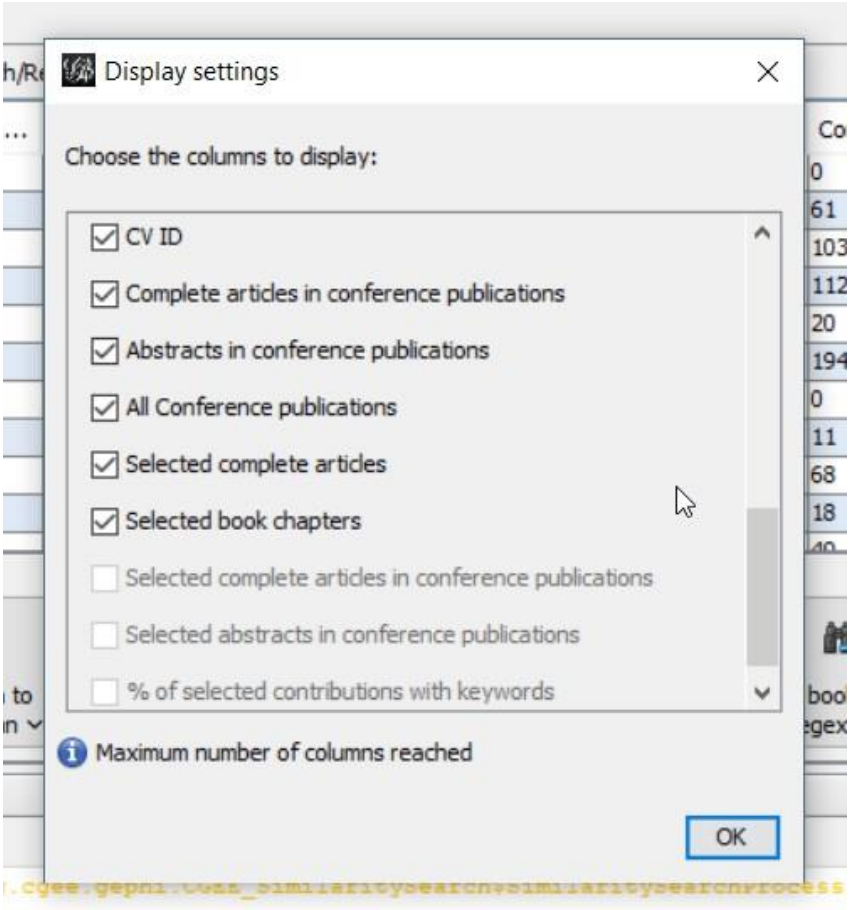


Figura 5.29: Seleção de colunas de atributos no laboratório de dados

5.4 Visualização e edição das contribuições Lattes

A rede de coautorias e de similaridade semântica é calculada a partir das contribuições bibliográficas **selecionadas** que constam nos currículos Lattes dos pesquisadores importados. Os detalhes dessas con- tribuições são visualizados na aba do Laboratório de Dados do Gephi:

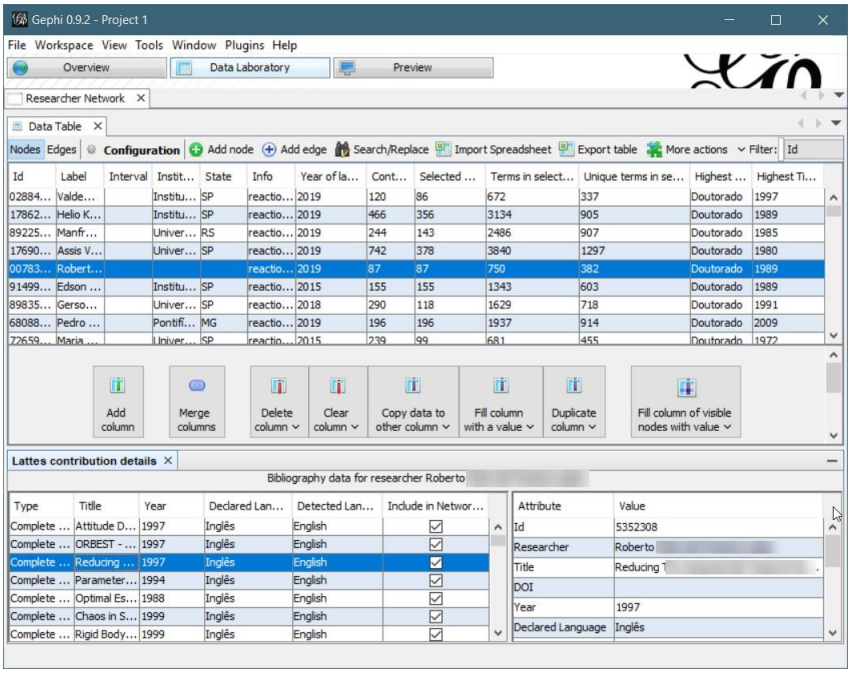


Figura 5.30: Exibição dos detalhes das contribuições bibliográfica de um Curriculo Lattes

O *CGEE Insight Net* permite a seleção de vários pesquisadores com as funcionalidades de *Shift-Click* e *Ctrl-Click* e mostra a produção para a união deles, exibindo adicionalmente uma coluna com os nomes. Se na lista das contribuiÇÕes for selecionad exatamente uma contribuição bibliográfica, os dados dela são exibidos no lado direito:

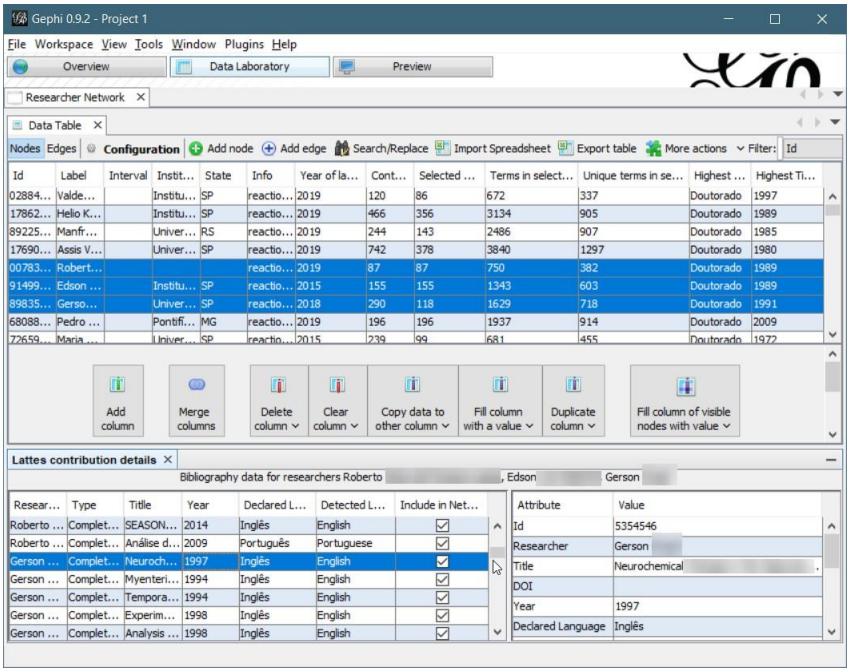


Figura 5.31: Exibição dos detalhes das contribuições bibliográfica de vários Curriculos Lattes

A tabela de contribuições fornece várias funcionalidades de edição e cópia.

- O idioma **declarado** das publicações individuais pode ser alterado de acordo com a lista de idio- mas que podem ser analisados.
- Contribuições individuais podem ser incluídas ou excluidas do escopo de cálculo da rede de co- autorias e similaridades semânticas, clicando na caixa de seleção na coluna “Include in Network build”. Neste caso, a seleção por critérios do diálogo *Build new Researcher Network* não estará disponível e exibirá apenas o botão `*Reset contribution selection criteria*`
- Uma lista de contribuições pode ser incluída ou excluída do escopo de cálculo da rede de co- autorias e similaridade semântica, selecionando as contribuições com *Shift-Click* ou *Ctrl-Click*. Depois deve ser clicado o botão direito do mouse em cima da tabela de contribuições. No menu *popup* que é exibido, o usuário pode selecionar *Include all selected contributions in network build* ou *Exclude all selected contributions from network build*:

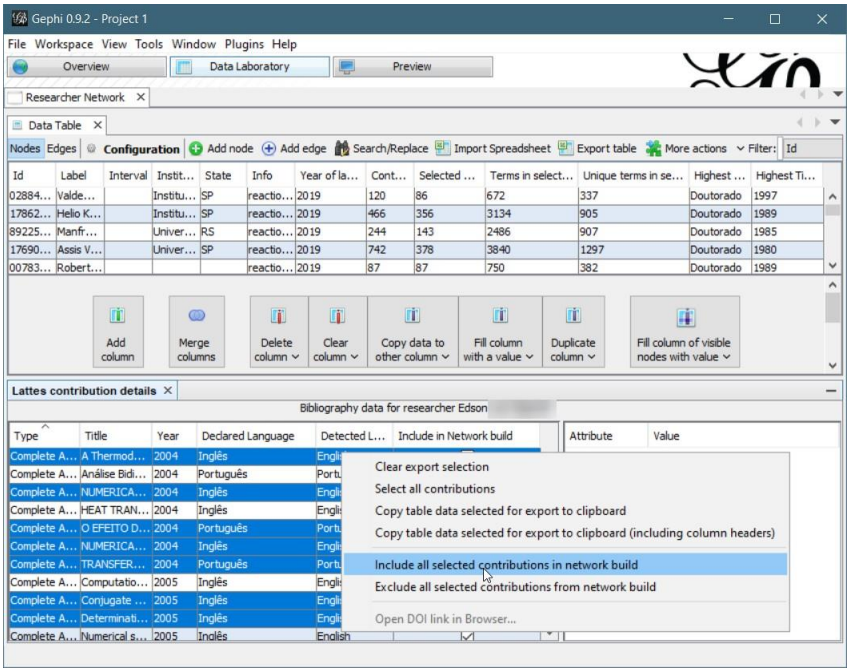


Figura 5.32: Inclusão ou exclusão de várias contribuições bibliográficas de Currículos Lattes no escopo do cálculo de rede

- Um clique com botão direito na lista de contribuições permite selecionar ou deselecionar todas as contribuições a partir das opções *Select all contributions* e *Clear export Selections*
- As contribuições selecionadas por *Shift-Click*, *Ctrl-Click* e pela opção *Select all contributions* podem ser exportadas para a área de transferência do computador e posteriormente inseridas em ferramentas como *Word*® ou *Excel*®. Para isso existem as opções *Copy table data selected for export to clipboard* e *Copy table data selected for export to clipboard (including column headers)* no menu de *popup* que é exibido com clique do botão direito na tabela de contribuições.
- Para contribuições que possuem um *Document Object Identifier (DOI)*, a referência bibliográfica pode ser aberta na internet com a opção *Open DOI link in Browser. . .*

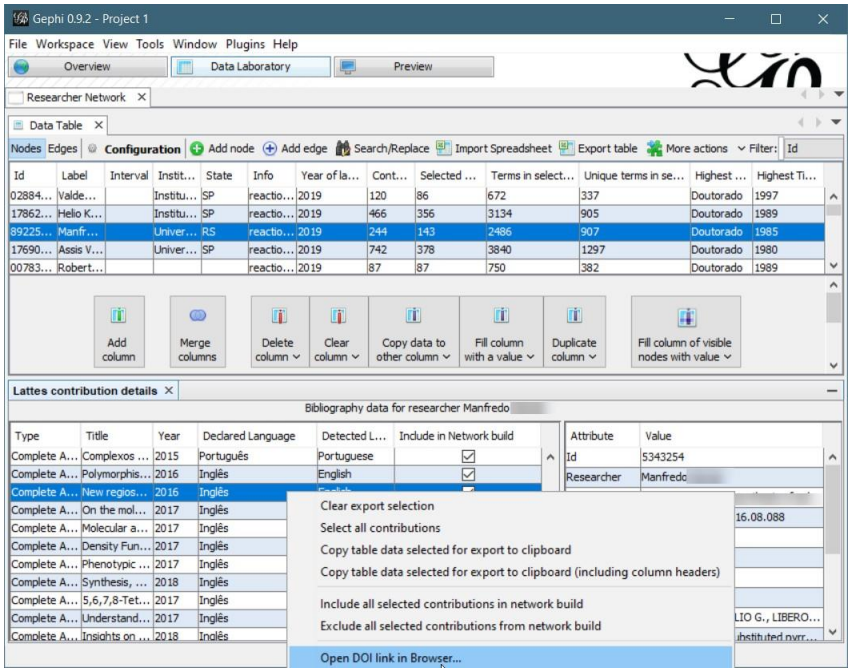


Figura 5.33: Exibir o Document Object Identifier no Browser

CAPÍTULO 6

Criação de redes de referências bibliográficas genéricas

O CGEE Insight Net oferece, com licença adicional, a criação de redes de referências bibliográficas a partir de arquivos dos serviços Web of Science® e Scopus®, bem como usando planilhas Excel® ou até qualquer outro formato estruturado de dados.

As redes bibliográficas são criadas a partir da similaridade semântica entre títulos e/ou resumos (*"abs- tracts"*) das publicações. A funcionalidade de criação de redes de co-ocorrências de palavras-chaves (ver Seção 7.5) complementa essa análise.

Para habilitar essa funcionalidade do *CGEE Insight Net*, a licença `INSIGHTNET_BIBLIOGRAPHY` deve ser instalada, conforme descrito na Seção 3.7. Essa licença é exibida assim no diálogo *Tools* > *Options* > *CGEE* > *License*:

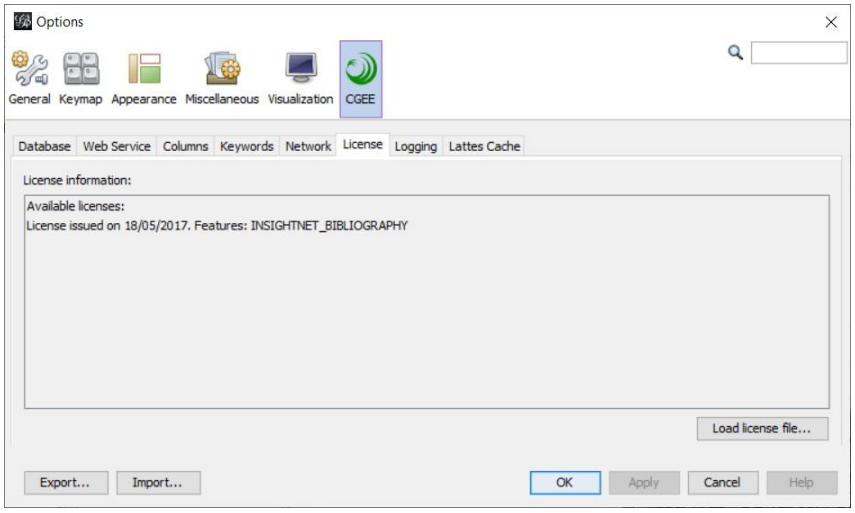


Figura 6.1: Licença requerida para o módulo de redes

bibliográficas Caso a licença esteja habilitada, aparece o sub-menu *"CGEE Insight Net Bibliography"*:

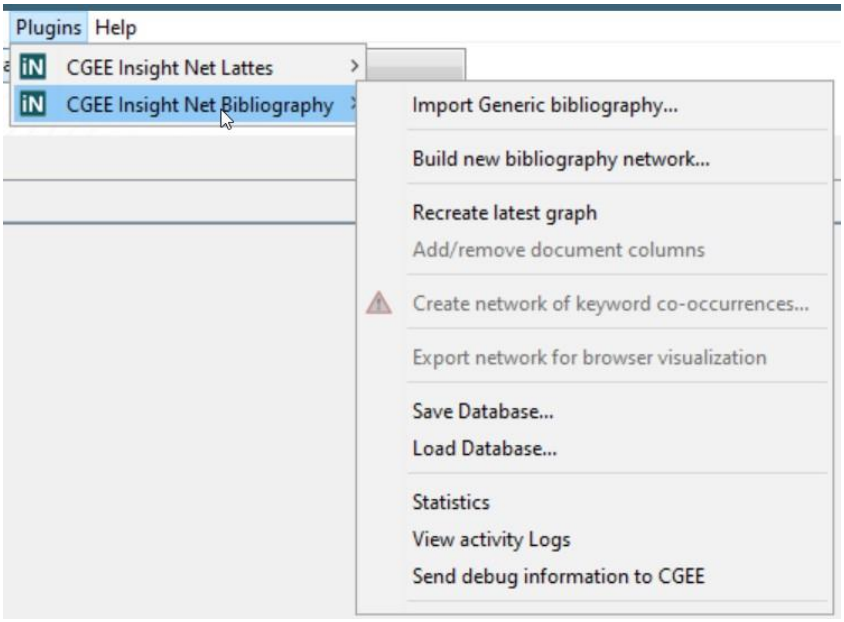


Figura 6.2: Sub-menu *CGEE Insight Net Bibliography*

6.1 Importação dos dados bibliográficos

Para importar dados bibliográficos, o usuário deve clicar em *Plugins* > *CGEE Insight Net Bibliography* > *Import Generic Bibliography* e escolher os arquivos a serem importados:

- Clicando em um arquivo, este será importado;
- Vários arquivos podem ser selecionados com *"Shift-Clique"* ou *"Ctrl-Clique"*, de acordo com os padrões de uso do sistema operacional;
- O usuário também pode selecionar um ou mais diretórios. Nesse caso, todos os arquivos nesse(s) diretório(s) serão importados.

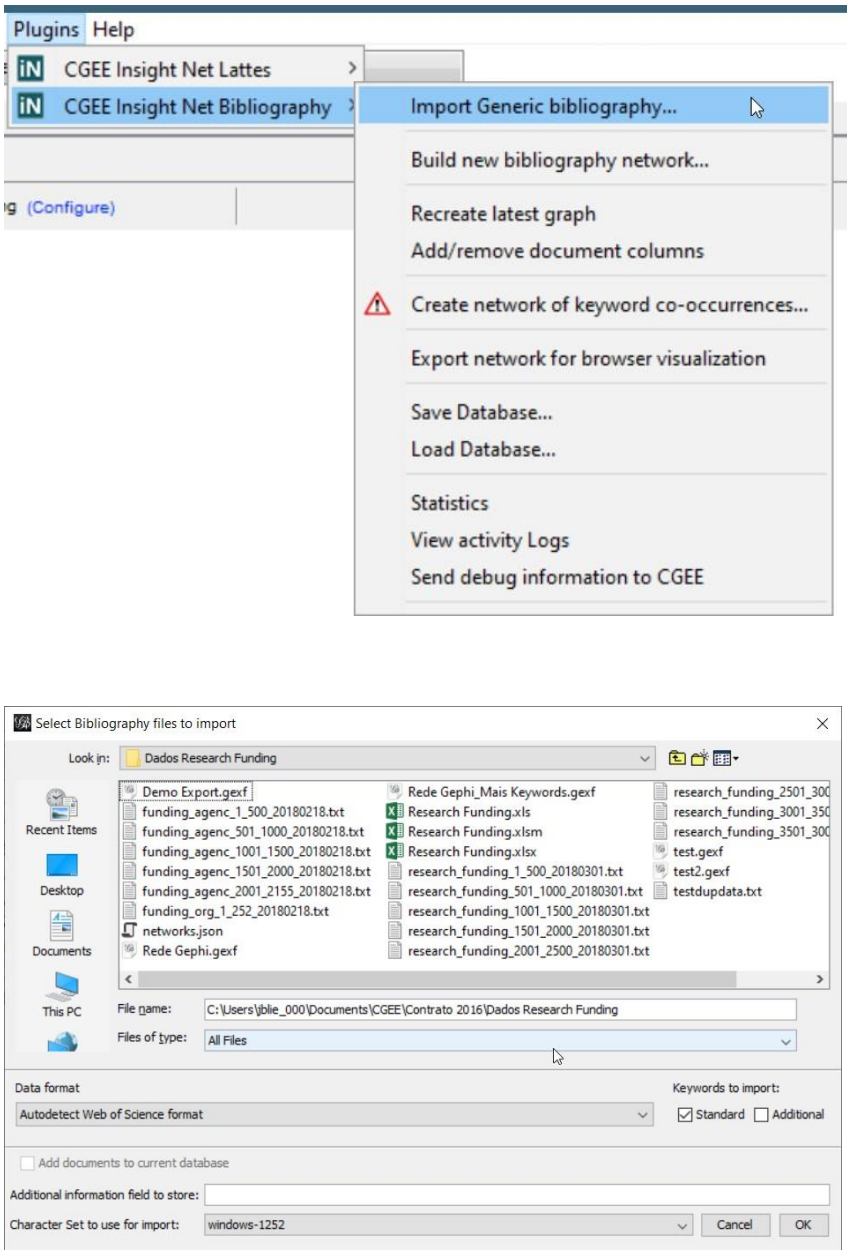


Figura 6.3: Importação de arquivos de bibliografia

6.1.1 Formato de dados

O módulo de bibliografia genérica traz uma lista de de formatos de dados prédefinidos, que atendem a demandas específicas. A escolha do formato apropriado na importação dos dados é essencial. Esta seção descreve os formatos disponíveis e orienta sobre os seus usos.

Formatos *Web of Science*® e *Scopus*®

Os serviços *Web of Science*® *Scopus*® disponibilizam dados em vários formatos que o *CGEE Insight Net* pode importar:

- *Comma separated values (CSV)*
- *Tab separated values (TSV)*
- *RIS* ou *Tagged*
- Dados genéricos em tabelas *Excel* (funcionalidade experimental)

O *Web of Science*® e o *Scopus* ainda permitem a disponibilização dos dados em formato "BibTeX", que pode ser importado pelo módulo específico do *CGEE Insight Net* (ver `bibtex`). Entretanto, o módulo de referências BibTeX carrega apenas um subconjunto das informações disponibilizadas.

A lista de formatos *Web of Science*® e *Scopus*® mostra, para cada um dos serviços as três opções men- cionadas acima, bem como a possibilidade de detectar o formato automaticamente, o que, geralmente, é a forma mais indicada:

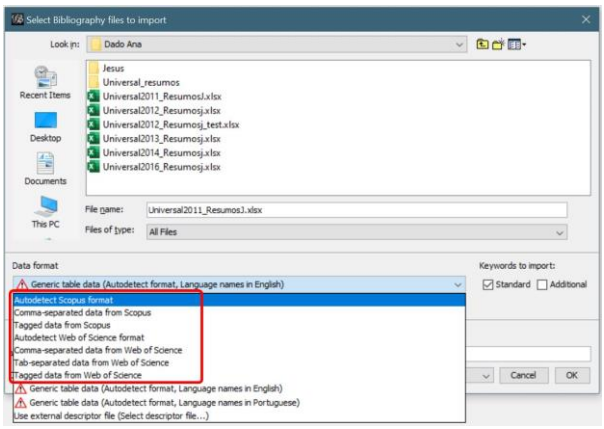


Figura 6.4: Formatos de arquivos de bibliografia dos serviços *Web of Science*® e *Scopus*®

As opções *Autodetect Scopus format* e *Autodetect Web of Science Format* permitem a importação de dados textuais desses serviços sem a necessidade de saber se o formato é *CSV*, *TSV* ou *Tagged*.

Arquivos no formato BibTeX devem ser importados pelo módulo de referências bibliográficas BibTeX (ver :numref:`bibtex`):

Formato "*Generic table data (Autodetect format)*"

O formato *Generic table data (Autodetect format)* usa planilhas Excel como arquivos de entrada. Neste caso, recomenda-se a importação de apenas um único arquivo, pois o usuário precisa especificar a aba da planilha a ser importada:

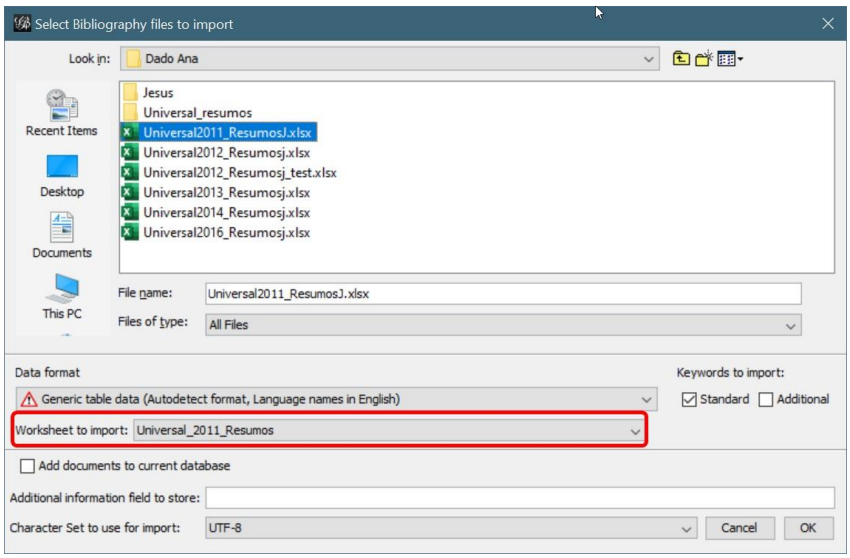


Figura 6.5: Seleção da aba da planilha excel no formato "*Generic Table data*

(Autodetect format)" A importação de dados genéricos presume os seguintes pré-

requisitos:

- Os dados a serem importados constam em uma única aba da planilha
- A primeira linha da planilha contém os cabeçalhos que descrevem o conteúdo das colunas
- A partir da segunda linha da planilha, cada linha contém exatamente uma referência bibliográfica
- O nome da coluna, como especificado na primeira linha, determina o comportamento da importa- ção, de acordo com a tabela em seguida. O nome deve ser escrito exatamente como especificado em seguida (observando letras maiúsculas e minúsculas) para obter o comportamento especifi- cado.
- As colunas adicionais, que possuem um nome que não consta na tabela em seguida serão impor- tadas como atributos dos nós no Gephi
- Colunas da planilha que não possuem nome na primeira linha são ignoradas

Nome da coluna	Comportamento da importação
Title	Título da publicação, usado para a criação da rede de similaridade semântica
Abstract	Resumo da publicação, usado para a criação da rede de similaridade semântica
Keywords	Contém a lista de palavras-chave da publicação, separadas por ponto-e-vírgula
ID	Contém um identificador único da publicação, usado para desduplicar referências bibliográficas durante a importação
DOI	Contém o *Document Object Identifier*, também usado para a desduplicação
Language	Contém o idioma da publicação.
Authors	Contem os autores da publicação, separados por ponto-e-vírgula
Year	Contém o ano da publicação, usado para filtrar as publicações na criação da rede de similaridade semântica
Document type	Contém o tipo de documento
Source title	Contém o título da revista, do livro ou do evento em que foi publicado o elemento

Formato "*Use external descriptor file*"

Para importar dados em formatos que não estão cobertos por uma das opções anteriores, o *CGEE In- sight Net* oferece com esta opção a possibilidade de especificar um formato em um arquivo externo. A especificação deste arquivo consta na `descriptors`.

6.1.2 Palavras-chave a serem importadas

Os serviços *Web of Science*® e *Scopus*® publicam as palavras-chave definidas pelo autor, bem como palavras-chave adicionais, extraídas pelas equipes das duas empresas a partir dos dados da publicação e do seu contexto. A opção *"Keywords to import"* permite a definição de quais delas serão importadas:

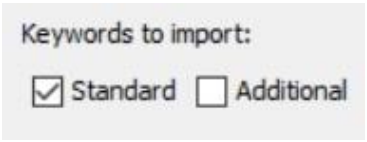


Figura 6.6: Seleção de palavras-chave a serem importadas

6.1.3 Apagar ou manter os dados do banco antes da importação

A opção *"Add documents to current database"* diferencia entre uma importação inicial e uma importa- ção incremental.

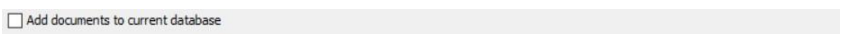


Figura 6.7: Opção de importação inicial ou incremental

Se essa opção for selecionada, as referências bibliográficas importadas serão acrescentadas às infor- mações já existentes na base. Caso uma referência bibliográfica já exista na base, a versão importada complementa as informações existentes.

Se a opção não for selecionada, todos os dados que já existem no banco de dados serão apagados antes da importação. Desta forma, os dados importados substituem os dados atuais.

6.1.4 Campo adicional de informação

Cada referência bibliográfica importada é representada como um nó no grafo criado. Esses nós possuem atributos, tais como o identificador DOI (*Digital Object Identifier* – vide http://www.doi. org/) da pu- blicação (atributo "*DOI*"), o seu título e outros. O atributo *"info"* dos nós é preenchido com o valor especificado no campo *"Additional information field to store"* durante a importação.

Figura 6.8: Campo adicional de informação

6.1.5 Conjunto de caracteres

A última opção do diálogo especifica o conjunto de caracteres (characterset) do(s) arquivo(s) a ser(em) importado(s). Contrário aos arquivos de Currículos Lattes, essa informação não está contida dentro de arquivos BibTex e precisa ser fornecida pelo usuário.

Figura 6.9: Definição do conjunto de caracteres

Os charactersets mais usados são "Windows-1252", para arquivos nativos do Windows e "UTF-8", para arquivos que foram baixados da internet. A seleção do *characterset* errado se manifesta em erros nas letras acentuadas durante a visualização do título e do resumo da referência bibliográfica.

6.1.6 Processo de importação

Durante a importação, o *CGEE Insight Net* mostra uma barra de progresso. Recomenda-se verificar a quantidade de referências bibliográficas no final da importação a partir da estatística do banco de dados (ver Seção 8.3).

Adicionalmente, o protocolo de execução (ver Seção 8.4) registra informações sobre o andamento da importação, de acordo com o grau de detalhe especificado na tela de configuração (ver Seção 3).

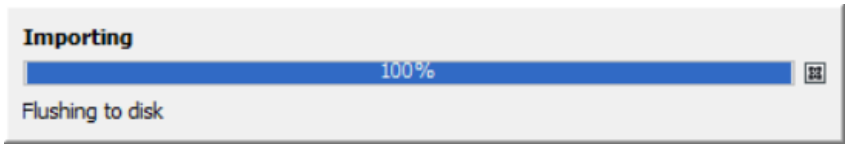


Figura 6.10: Importação de referências bibliográficas

6.2 Formação da rede

Depois da importação das referências bibliográficas na base de dados, a rede é formada a partir das pes- quisas por similaridade contextual. Os passos 2-4 da Tabela 4.1 são realizados em uma única operação, transformando o conteúdo do banco de dados em um grafo.

Para formar a rede, o usuário deve clicar em *Plugins* > *CGEE Insight Net Bibliography* > *Build new bibliography network* e preencher ou confirmar os dados do diálogo que é exibido:

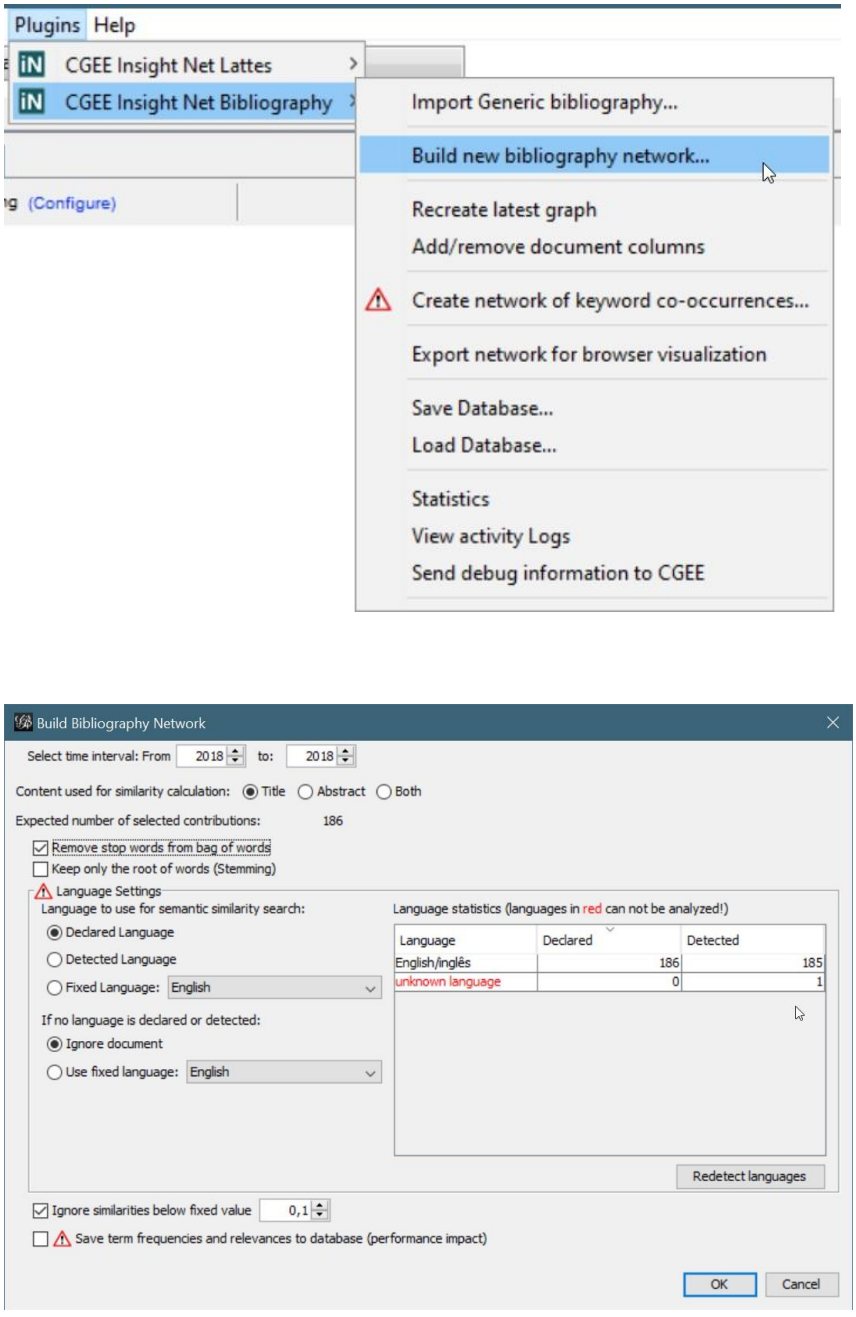


Figura 6.11: Menu e diálogo da formação da rede de referências

bibliográficas As opções do diálogo são explicadas em seguida.

6.2.1 Escopo da rede formada

Na parte superior do diálogo o usuário especifica quais publicações farão parte do escopo da formação da rede. No módulo de referências bibliográficas genéricas, o único critério é o intervalo de anos de publicação.

Destaca-se novamente que a rede de referências bibliográficas será montada apenas para as contribuições selecionadas.

6.2.2 Conteúdo considerado para o cálculo de similaridade contextual

A parte inferior do diálogo permite a seleção das opções da pesquisa por similaridade semântica.

- A seleção "*Content used for similarity calculation*" determina se a pesquisa por similaridade se- mântica considera apenas o título da referência bibliográfica ("*Title*"), apenas o resumo ("*Abs- tract*") ou ambos ("*Both*").
- O usuário pode selecionar se os pré-processamentos dos termos "*Stop words*" e "*Stemming*" serão realizados ou não. Esses dois algoritmos dependem da definição correta do idioma da referência bibliográfica.
  - *Stop Words* são as palavras mais frequentes de cada idioma, que não agregam informação aos termos identificados e serão eliminados da pesquisa. Os stop words são implementados apenas para as referências bibliográficas em Inglês e Português.
  - O *Stemming* reduz, em um algoritmo específico por idioma, cada palavra a uma raiz que desconsidera flexões gramaticais. Nesse momento, apenas os idiomas Português e Inglês são tratados pelo stemming. Referências bibliográficas em outros idiomas permanecem na forma original.
  - Se qualquer uma dessas opções for selecionada, o idioma do texto se torna relevante. Neste caso, aparece no diálogo a estatística de idiomas declarados e detectados, de acordo com a seleção no campo "*Context used for similiarity calculation*".

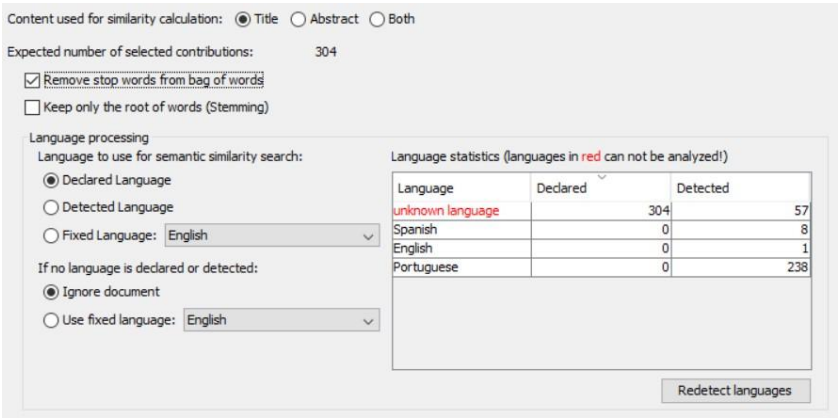


Figura 6.12: Estatística de idiomas detectados e declarados

  - O botão "*Redetect languages*" permite realizar uma nova detecção de idiomas com

parâme- tros diferentes daqueles configurados na tela "*Languages*" da configuração do plugin (ver Seção 3.6):

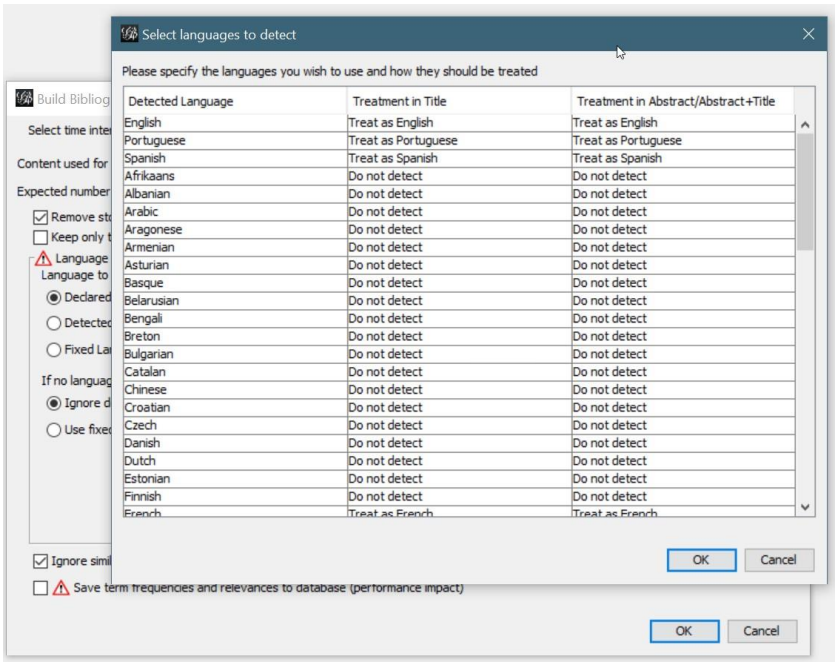


Figura 6.13: Nova detecção de idiomas

– Adicionalmente, o usuário pode alterar os idiomas declarados e detectados no laboratório de dados do *Gephi*, para corrigir possíveis erros nesses dados:

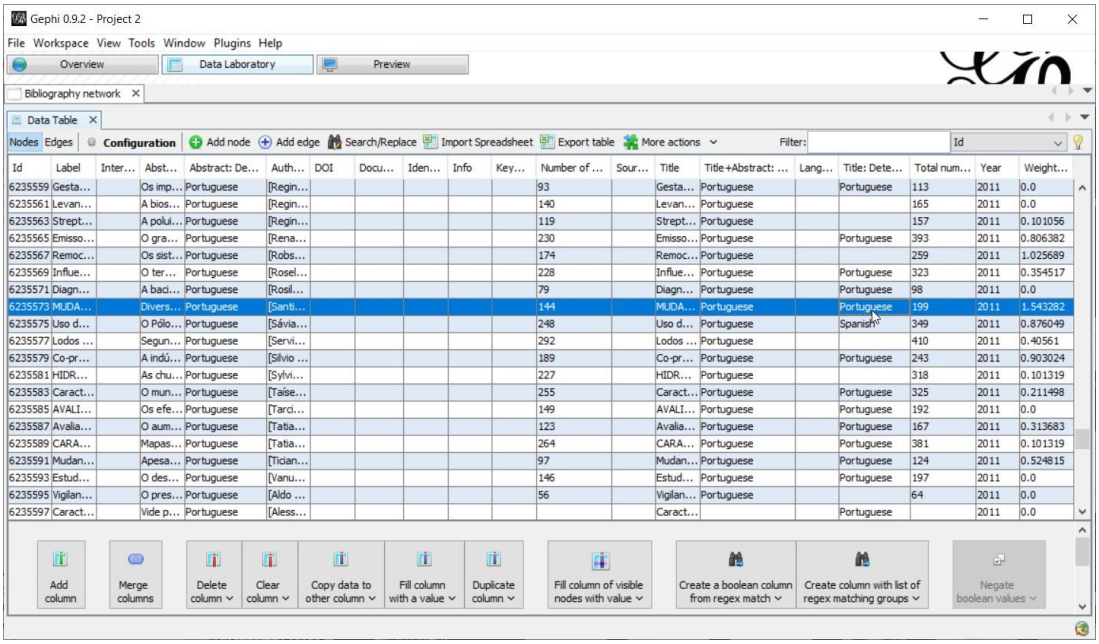


Figura 6.14: Correção manual dos idiomas das contribuições no laboratório de dados do Gephi

• O cálculo da similaridade semântica gera arestas entre praticamente qualquer par de referências bibliográficas, a maioria com baixos valores de similaridade, que não agregam informações rele- vantes ao conteúdo do gráfico. Por esse motivo, existem três métodos para reduzir a quantidades de arestas na rede:

– "*Ignore similarities below fixed value*": valores abaixo de um limite especificado podem ser

desconsiderados, produzindo o valor final zero como similaridade semântica.

- "*Sparsify network automatically*": Um algoritmo automático [7] ainda experimental é utili- zado para reduzir a quantidade de arestas na rede.
- Para o cálculo de similaridade podem ser considerados apenas os termos mais relevantes das contribuições. Esta configuração é feita no diálogo de opções do CGEE Insight Net (ver Seção 3). Se um percentil de relevância dos termos for definido, uma mensagem correspon- dente é exibida:

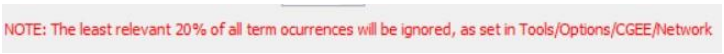


Figura 6.15: Aviso sobre configuração de limite inferior de relevância

CAPÍTULO 7

Análise das redes criadas

7.1 Filtragem dos resultados

O *CGEE Insight Net* define quatro filtros que modificam a exibição do grafo, eliminando informações específicas determinadas pelo usuário. Esses filtros são exibidos na aba "*Filters*" do Gephi, na categoria "CGEE":

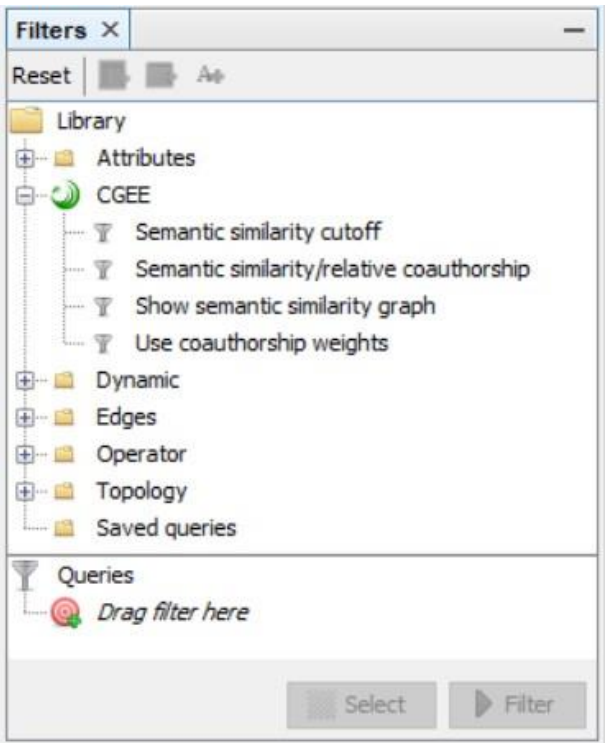


Figura 7.1: Filtros da categoria "CGEE"

Alguns desses filtros são relevantes apenas para redes que possuem arestas de coautoria e também de similaridade semântica. Outros podem ser usados em ambos os tipos de redes.

Para escolher um dos filtros, o mesmo deve ser selecionado pelo usuário com clique duplo. Existem dois tipos de aplicar o filtro:

1. O filtro é aplicado com um clique no botão "*Filter*". Desta forma, a rede muda de acordo com o filtro selecionado.
2. Existe ainda o botão "*Select*", que **destaca** os elementos da rede que atendem ao critério do filtro. Entretanto, este funcionalidade é usado, principalmente, com os filtros padrão do Gephi. Para os filtros específicos do CGEE, este botão não agrega valor.

7.1.1 Filtro "*Show semantic similarity graph*"

Esse filtro não possui parâmetros de configuração e tem efeito apenas em redes que possuem ambos os tipos de arestas (coautoria e similaridade semântica). Ele elimina todas as arestas verdes e transforma as arestas pretas em vermelhas ao exibir o grafo. Como peso das arestas, nesse caso, é usada exclusivamente a similaridade semântica.

7.1.2 Filtro "*Semantic similarity/relative coauthorship*"

Esse filtro também traz resultados apenas em redes que possuem arestas dos dois tipos (coautoria e similaridade semântica). Ele permite a definição de um intervalo de exibição das arestas, controlado pelo valor do atributo "*Semantic similarity/relative coauthorship*". Arestas que representam colaborações baseadas na coautoria com similaridade semântica zero ("arestas verdes") sempre carregam o valor zero nesse atributo. Arestas com zero coautorias, que apenas possuem similaridade semântica ("arestas vermelhas") são consideradas com um valor alto, além do valor de qualquer aresta preta. O intervalo de exibição pode ser escolhido com dois marcadores. O gráfico abaixo desses marcadores representa um histograma de valores encontrados no grafo:

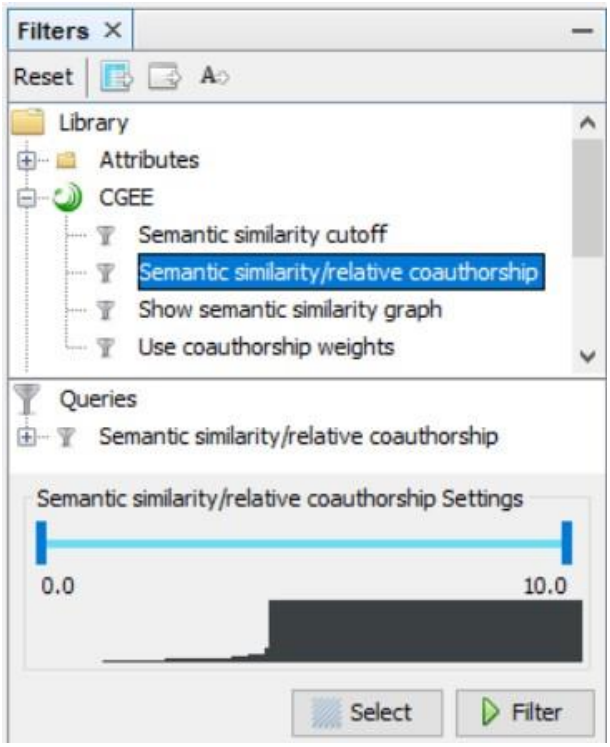


Figura 7.2: Parâmetros do filtro *Semantic similarity/relative coauthorship*

7.1.3 Filtro "Semantic similarity cutoff"

Esse filtro funciona em redes que possuem arestas do tipo "similaridade semântica" e desconsidera na exibição qualquer similaridade semântica abaixo do limite especificado pelo usuário.

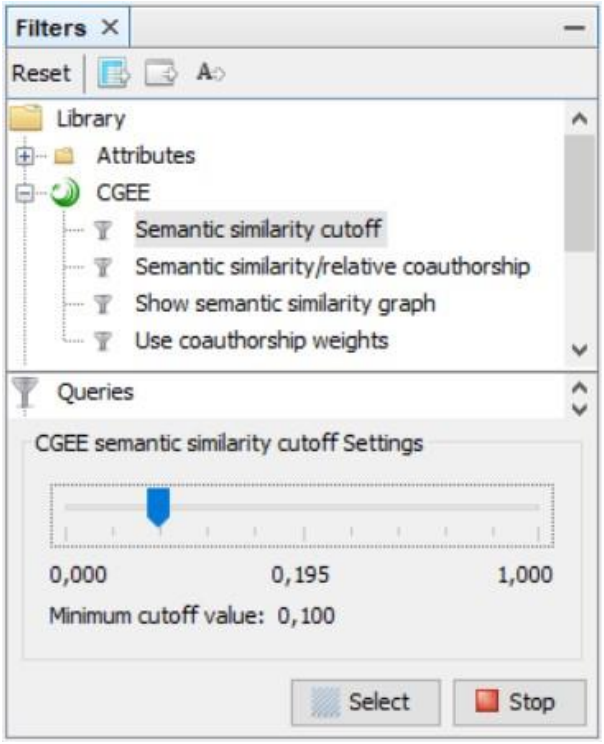
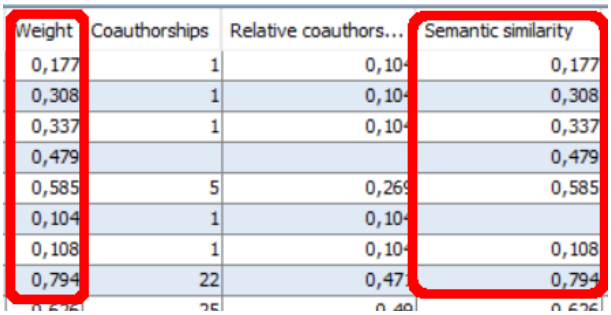


Figura 7.3: Parâmetro do filtro *Semantic similarity cutoff*

Aumentando o valor mínimo da similaridade semântica para o valor máximo possível (1.0), todas as arestas de similaridade semântica serão eliminadas e o grafo exibirá apenas as arestas de coautorias, caso existam.

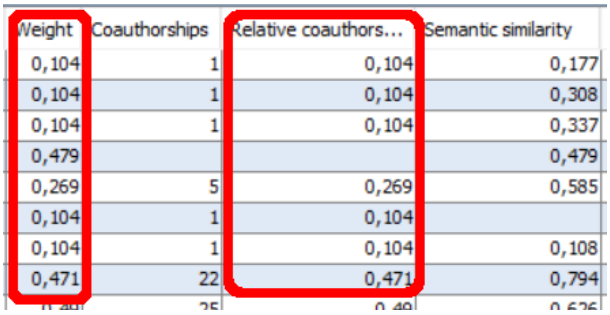
7.1.4 Filtro "*Use coauthorship weights*"

Em grafos que possuem arestas com valores de similaridade contextual bem como de coautorias ("arestas pretas"), normalmente se usa como peso a similaridade contextual/semântica. O filtro "*Use coauthorship weights*" permite, nesses casos, a aplicação das coautorias relativas como peso da aresta, conforme demonstram os seguintes diagramas:



Weight	Coauthorships	Relative coauthors...	Semantic similarity
0,177	1	0,10	0,177
0,308	1	0,10	0,308
0,337	1	0,10	0,337
0,479			0,479
0,585	5	0,26	0,585
0,104	1	0,10	
0,108	1	0,10	0,108
0,794	22	0,47	0,794

Figura 7.4: Sem filtro "*Use coauthorship weights*"



Weight	Coauthorships	Relative coauthors...	Semantic similarity
0,104	1	0,104	0,177
0,104	1	0,104	0,308
0,104	1	0,104	0,337
0,479			0,479
0,269	5	0,269	0,585
0,104	1	0,104	
0,104	1	0,104	0,108
0,471	22	0,471	0,794

Figura 7.5: Com filtro "*Use coauthorship weights*"

7.2 Análise de clusters

As funcionalidades do Gephi e os *plug-ins* existentes permitem a determinação de "*clusters*", grupos de pesquisadores que entre si possuem mais ligações do que com pesquisadores externos. O algoritmo mais utilizado no Gephi bem é representado pela estatística "*Modularity*".

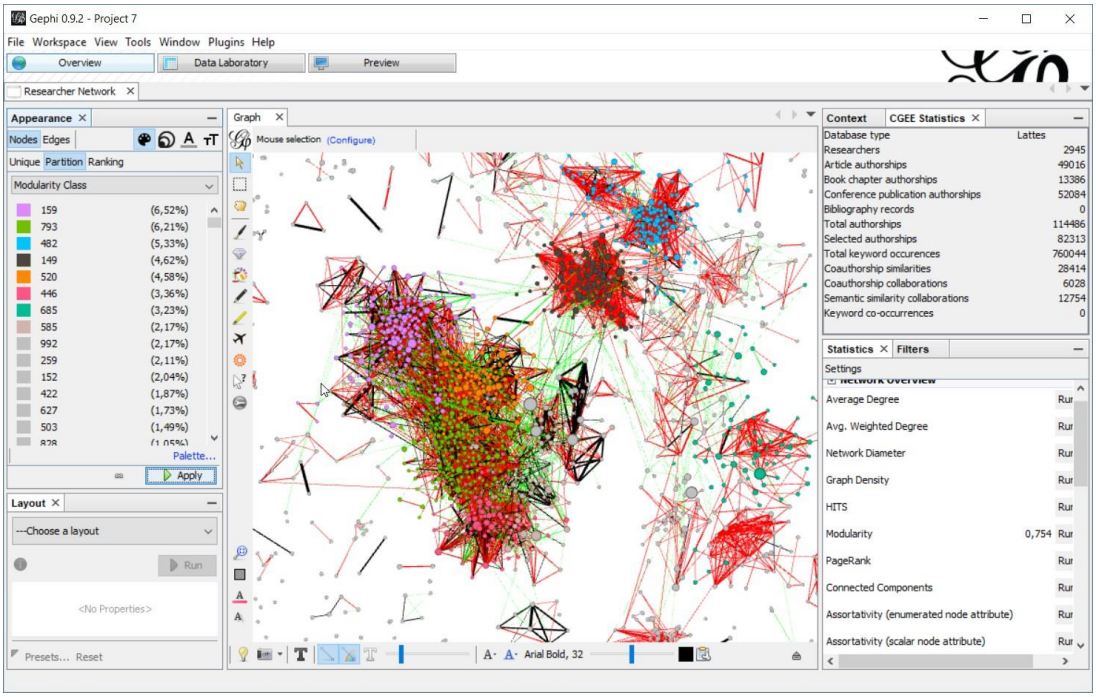


Figura 7.6: Clusters na rede gerada

7.3 Análise de assortatividade

A assortatividade (também conhecida como “homofilia”) de uma rede descreve a tendência da existência de uma aresta entre dois nós que possuem valores semelhantes em um atributo selecionado [Assort].

Por exemplo, em redes sociais, existem tendências fortes de estabelecer amizades dentro do mesmo nível educacional ou da mesma nacionalidade. Nesses casos, a assortatividade de uma rede social teria um valor alto positivo com relação aos atributos “nível educacional” ou “nacionalidade”.

Já em redes de relacionamento sexual, a preferência geralmente é pelo gênero oposto, levando a assor- tatividade com relação ao atributo “gênero” para um valor negativo.

Em redes não-assortativas, a existência da aresta não é correlacionada ao atributo selecionado e o valor da assortatividade em relação ao atributo escolhido é próximo de zero.

O *CGEE Insight Net* permite a análise da assortatividade em relação a um atributo selecionado a partir de duas estatísticas do Gephi no painel “*Statistics*”:

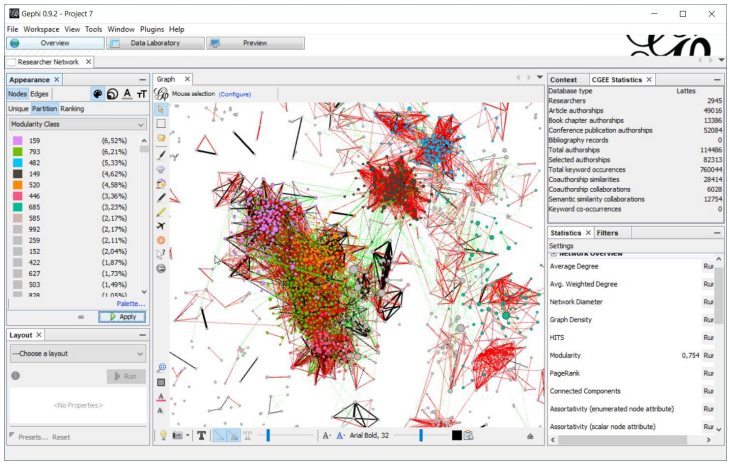
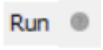


Figura 7.7: Estatísticas de

Assortatividade Dependendo do tipo de atributo, uma das duas

estatísticas deve ser escolhida

- Para atributos enumerados (ou categóricos) deve ser escolhida a estatística “*Assortativity (enume- rated node attribute)*”. Atributos enumerados são aqueles em que os valores não possuem ordem, tais como valores textuais ou categorias numeradas.
- Atributos numéricos que possuem uma ordem entre si podem ser avaliados com a estatística “*As- sortativity (scalar node attribute)*”. São aqueles onde o valor representa uma medida numérica.

Para calcular a assortatividade, o usuário deve clicar no botão 

da estatística

selecionada e esco- lher o(s) atributo(s) com relação ao qual(is) a assortatividade deve ser calculada:

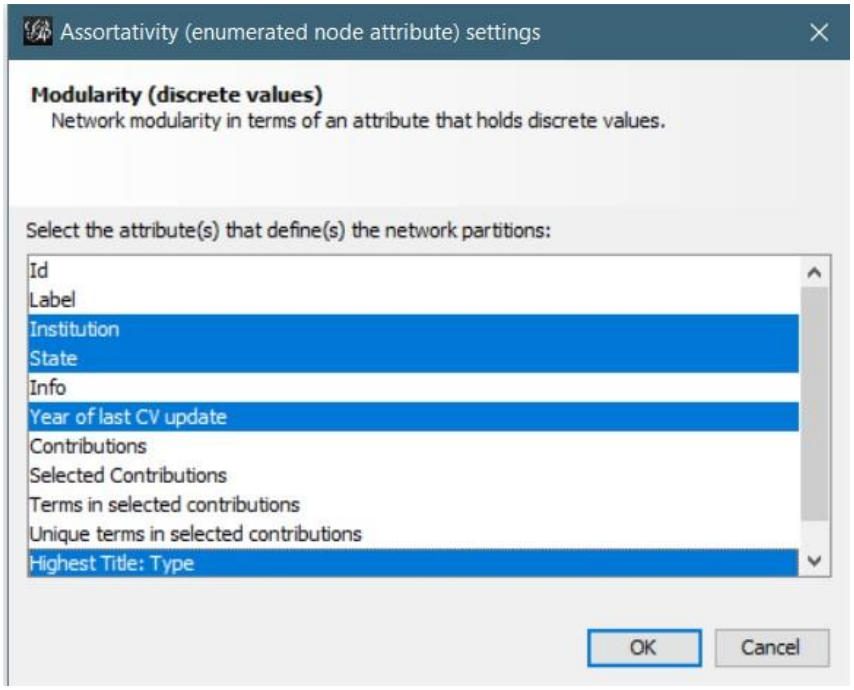
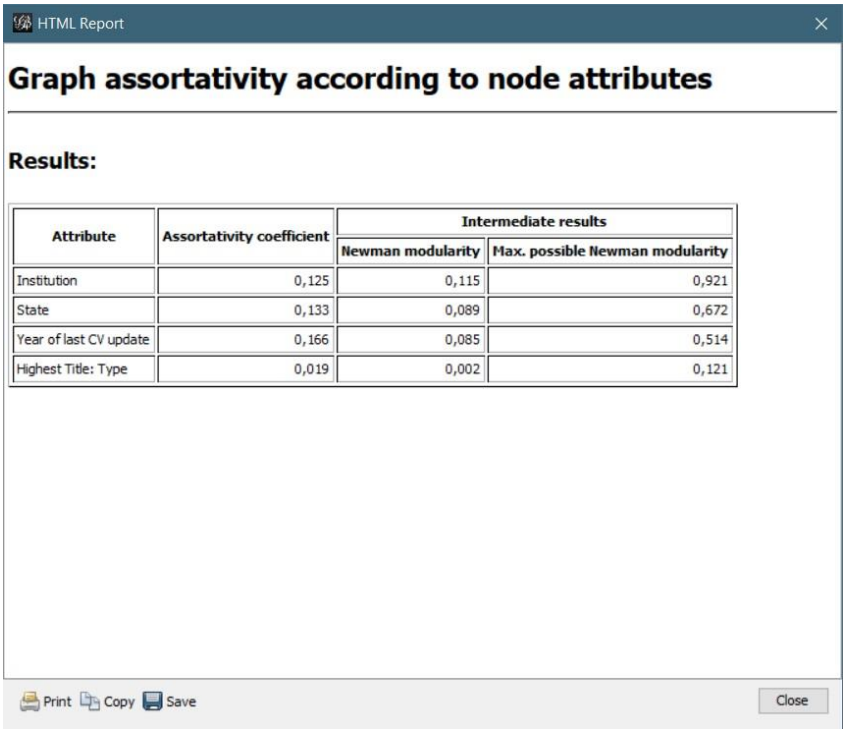


Figura 7.8: Seleção do(s) atributo(s) de assortatividade

Clicando em "*OK*", a assortatividade é calculada e aparece no diálogo de resultado, bem como na lista de cálculos estatísticos do Gephi:



Attribute	Assortativity coefficient	Intermediate results	
		Newman modularity	Max. possible Newman modularity
Institution	0,125	0,115	0,921
State	0,133	0,089	0,672
Year of last CV update	0,166	0,085	0,514
Highest Title: Type	0,019	0,002	0,121

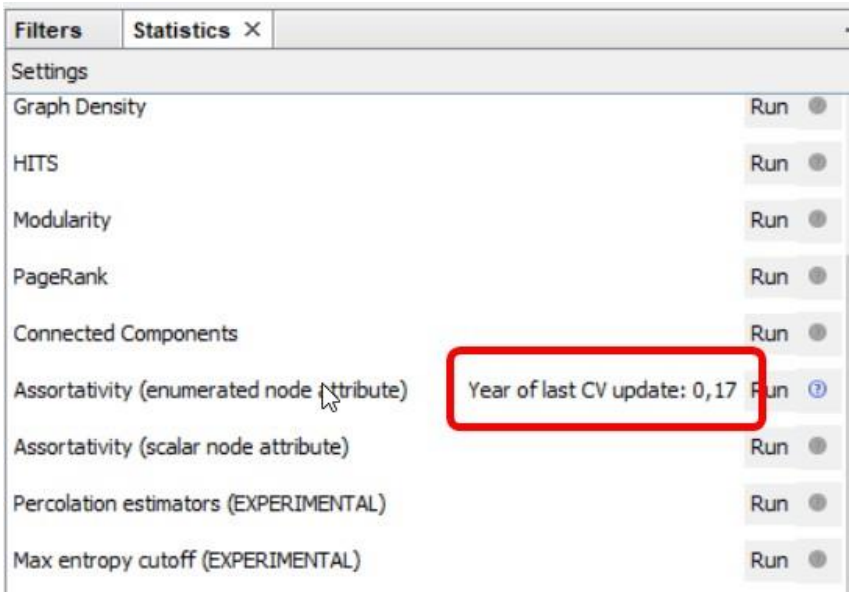


Figura 7.9: Resultados do cálculo de assortatividade

7.3.1 Estimador de percolação

Essa funcionalidade avançada é atualmente usada apenas para testes internos do CGEE. Os fundamentos matemáticos podem ser encontrados na literatura especializada [Percolation].

7.4 Análise das palavras-chave

Para aprofundar a análise das redes, as palavras-chave dos Currículos Lattes dos pesquisadores e das contribuições são importadas no banco de dados. Para redes bibliográficas, as palavras-chave são extraí- das dos arquivos carregados.

Essas palavras-chave permitem, quando examinadas em conjunto, a identificação das áreas de trabalho dos pesquisadores representados pelo conjunto de seus currículos.

Para redes de referências bibliográficas, as palavras-chave caracterizam o conteúdo da publicação e permitem, em conjunto, uma estimativa geral dos conteúdos das publicações.

O CGEE Insight Net permite a visualização dessas palavras-chave, que constam **apenas** no seu banco de dados e não no grafo do Gephi. Desta forma, é essencial que o banco de dados seja **coerente** com o grafo. Incoerências podem surgir se um grafo Gephi for carregado por um arquivo ".gephi" ou ".gexf" e se o banco de dados não possuir o mesmo conteúdo que esse arquivo. Neste caso, sugere-se nova importação dos Currículos Lattes ou das referências bibliográficas no banco de dados, nova geração do grafo ou a recuperação do grafo a partir do banco de dados, conforme descrito na Seção 8.1.

Para realizar a pesquisa de palavras-chave, o usuário deve clicar no símbolo  na barra lateral da tela "*Graph*" do Gephi:

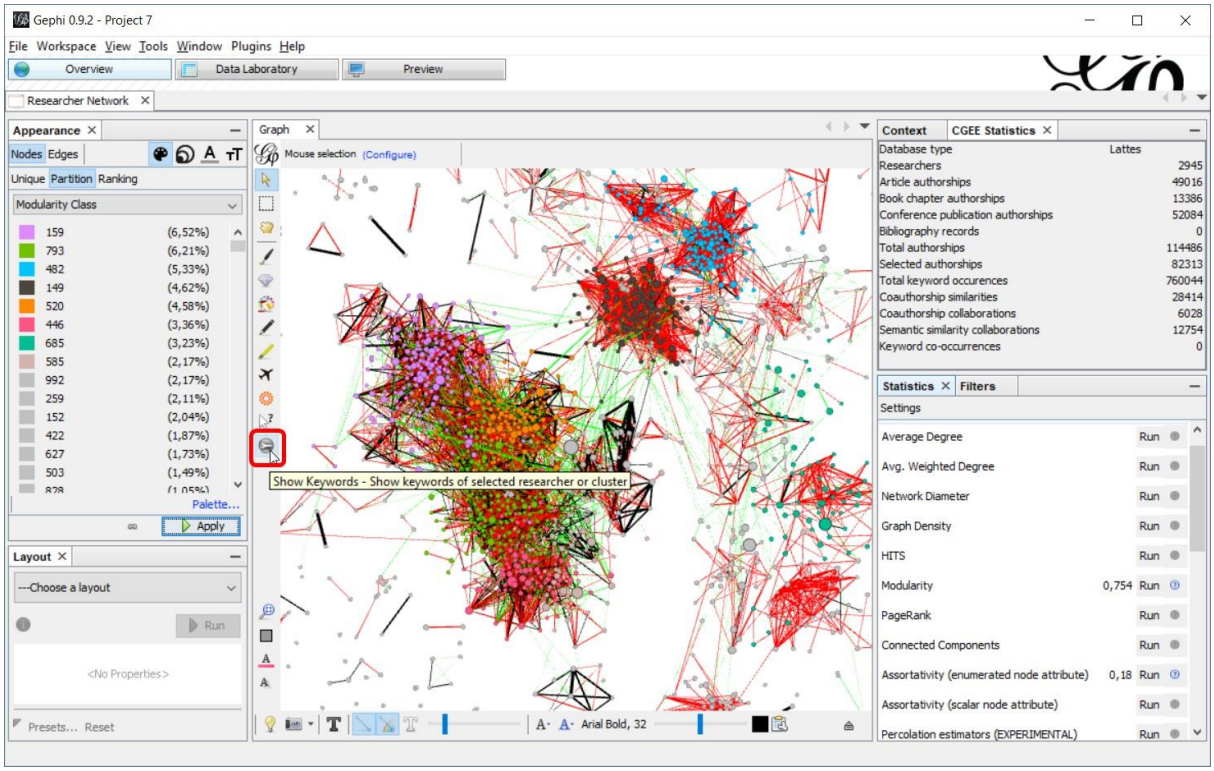


Figura 7.10: Seleção da funcionalidade "Palavras-chave"

Clicando nesse símbolo e selecionando nós no grafo, aparece a janela "*CGEE Keywords*", que mostra as palavras-chave do nó selecionado e as frequências (quantidades de ocorrências) de cada uma:

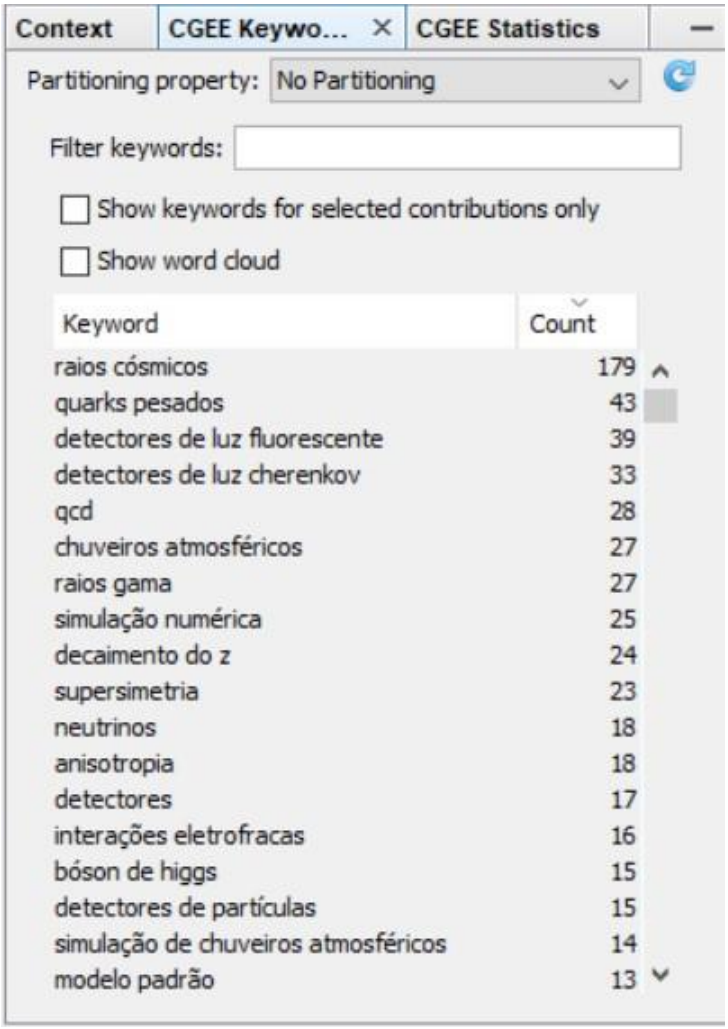


Figura 7.11: Janela de palavras-chave com

frequências Essa janela permite algumas configurações que serão

explicadas em seguida.

7.4.1 Filtragem das palavras-chave

A lista de palavras-chave pode ser filtrada para exibir apenas alguns dos resultados. Existem dois tipos de filtros, descritos em seguida. Em ambos os casos, o texto exibido na lista muda de cor e aparece na tela o número de palavras-chave que são eliminadas da lista pelo filtro.

Filtragem por termo digitado

O usuário pode digitar um texto na caixa em cima da lista. Nesse caso, apenas as palavras-chave que contêm o texto digitado aparecem na lista:

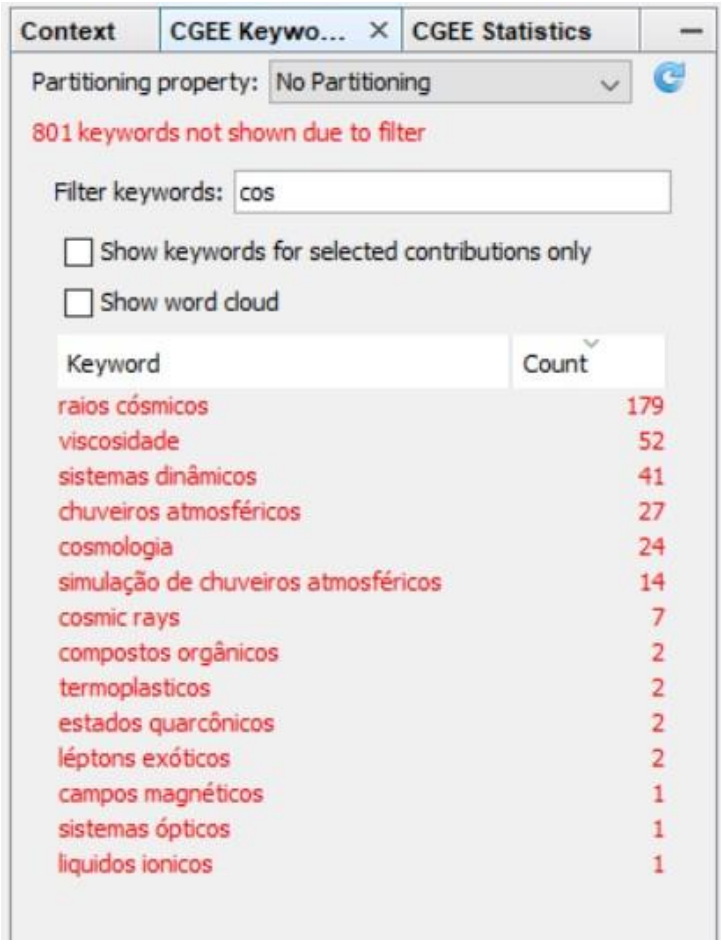


Figura 7.12: Filtragem de palavras-chave por termo

Filtragem por contribuições selecionadas

Essa função se aplica apenas em redes de Currículos Lattes, em que existem vários elementos que permitem especificar palavras-chave:

- Na descrição da formação,
- Nas atividades de pesquisa e desenvolvimento,
- Nas produções científi cas (artigos, capítulos de livros, trabalhos em eventos),
- Nas orientações de graduação, mestrado e doutorados
- Outros

A exibição de todas as palavras-chave de um único pesquisador permite uma visão global das áreas de atuação contribuindo parauma avaliação rápida do conteúdo semântico integrado de toda a produção do pesquisador. Por outro lado, considerando que a identificação das coautorias, da similaridade contex- tual e o agrupamento em clusters utilizam apenas as contribuições selecionadas durante a pesquisa de similaridade, é razoável exibir apenas as palavras-chave dessas contribuições selecionadas.

A opção "*Show only keywords for selected contributions*" permite alternar entre as duas formas de exibição descritas:

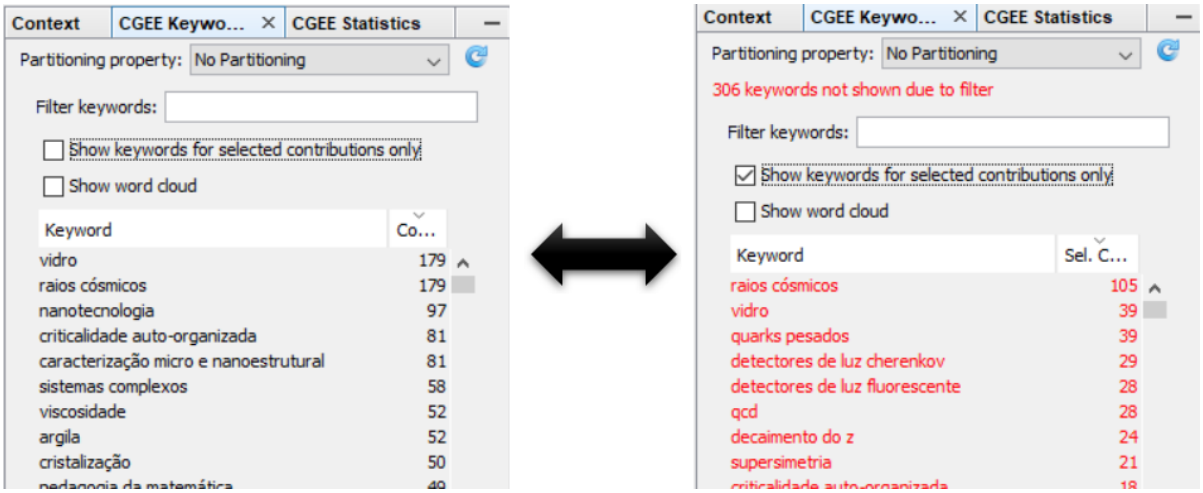


Figura 7.13: Filtragem das palavras-chave pelas contribuições selecionadas

7.4.2 Palavras-chave por nó ou por cluster de nós

A lista "*Partitioning property*" na parte superior permite selecionar se a janela exibe apenas as palavras- chave do nó selecionado ("*No clustering*") ou as palavras-chave de todos os nós de um grupo definido (partição). Para determinar a partição, um atributo numérico, ou os atributos "Info" ou "*Institution*", pré- definidos nos nós, devem possuir o mesmo valor para todos os membros do grupo. Casos particulares de partições importantes são os agrupamentos calculados por diferentes métodos no Gephi. Nesses casos, a partição é definida a partir das arestas entre os nós, sendo, portanto um atributo pós-processado. Os algoritmos de agrupamento (*clustering*) do Gephi descritos na Seção 7.2 usam atributos diferentes para especificar o número do cluster e o usuário precisa selecionar aquele mais adequado à sua análise.

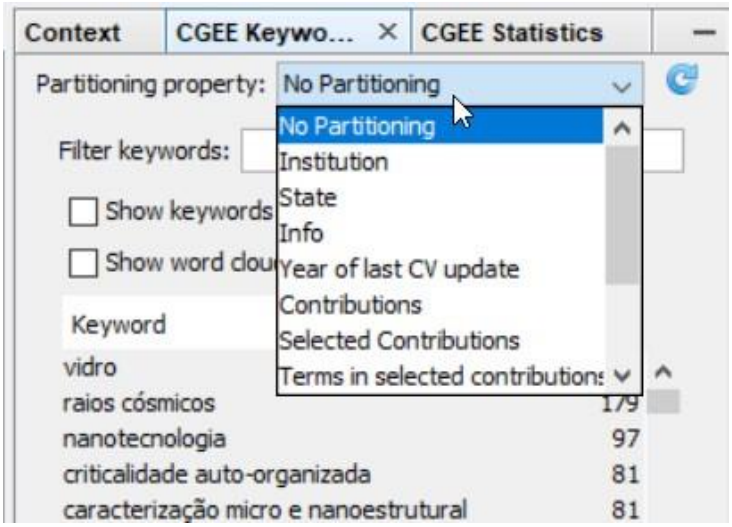


Figura 7.14: Seleção do atributo que define a partição dos nós

Depois de executar um algoritmo de *clustering*, a lista precisa ser atualizada manualmente,

clicando no símbolo  . Além de todos os valores numéricos integrais, a lista de possíveis atributos de particiona- mento mostra os seguintes atributos em redes de Currículos Lattes:

- Todos os atributos numéricos integrais

- *Info*
- *Institution*
- *State*
- *Gender*
- Todas as informações sobre a formação dos pesquisadores

Alguns dos valores numéricos não representam clusters, no sentido de agrupamento, como por exemplo o número de contribuições, mas que ainda são partições.

Se a opção "*Partitioning property*" for ativada, a lista mostra todas as palavras-chave da partição da qual o nó selecionado pertence, junto com sua quantidade de nós. Para destacar o fato que a lista mostra as palavras-chave de uma partição e não do nó individual, as palavras-chave são exibidas em verde. Dependendo da configuração (ver Seção 3.4), a lista mostra ainda a porcentagem ou a quantidade de nós em que cada palavra-chave é encontrada:

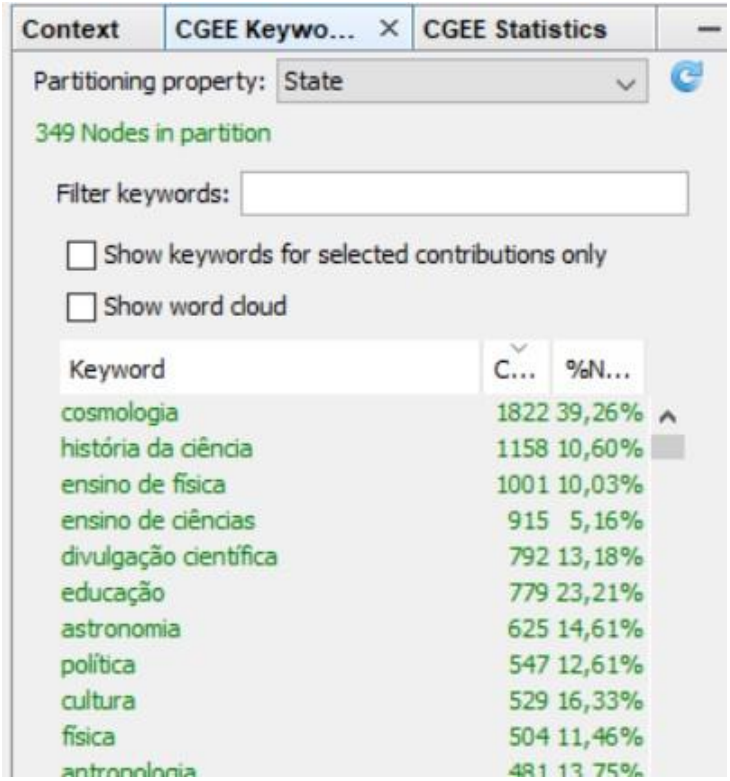


Figura 7.15: Palavras-chave de um cluster de 349 nós

Nesse exemplo, a palavra-chave mais frequente ("cosmologia") tem 1.822 ocorrências em 137 nós (39,26% de 349 nós).

A filtragem dessa lista de palavras-chave por partição, de acordo com a Seção 7.4.1 exibe as palavras- chave na cor laranja. Observe-se que no exemplo em baixo a lista foi configurada para mostrar a quan- tidade absoluta de nós (ver Seção 3.4):

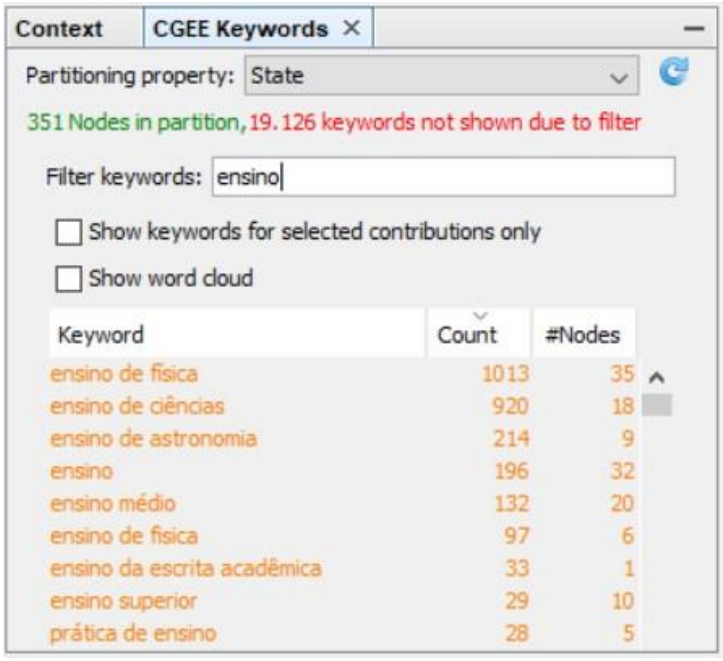


Figura 7.16: Lista de palavras-chave por partição, filtrada

Conforme descrito na Seção 3.4, pode ser calculada a relevância das palavras-chave por partição. Caso a referida opção tenha sido selecionada na tela de configuração, sempre ressaltando tratar-se de uma funcionalidade experimental, a relevância aparece como coluna adicional na lista de palavras-chave:

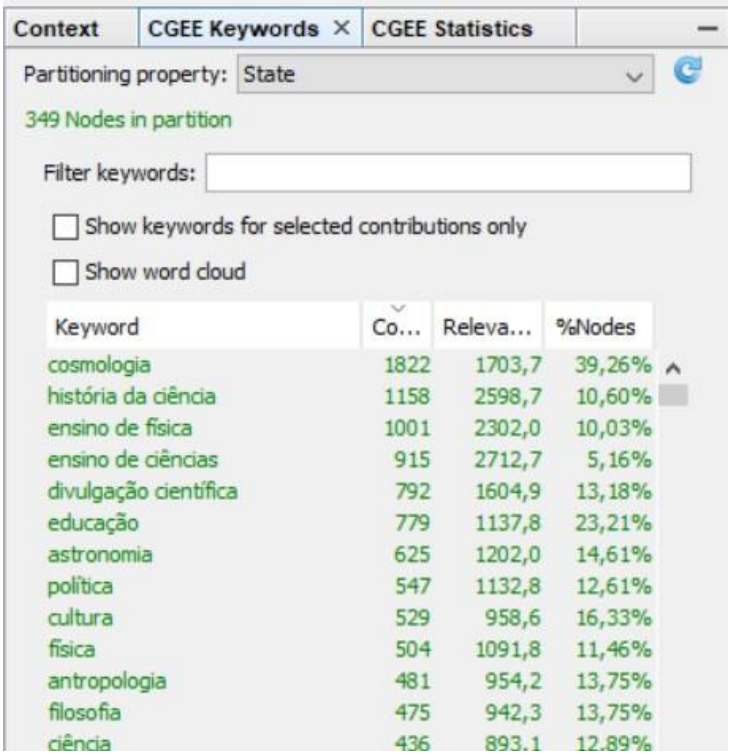


Figura 7.17: Relevância das palavras-chave exibidas na lista

7.4.3 Seleção de nós a partir das palavras-chave

Conforme descrito na Seção 7.4, a lista de palavras-chave é exibida a partir da seleção de um ou mais nós e o conteúdo mostrado depende da configuração. Como funcionalidade adicional, itens da lista de palavras-chave podem ser selecionados e, nesse caso, todos os nós que usaram esses itens são destacados no grafo:

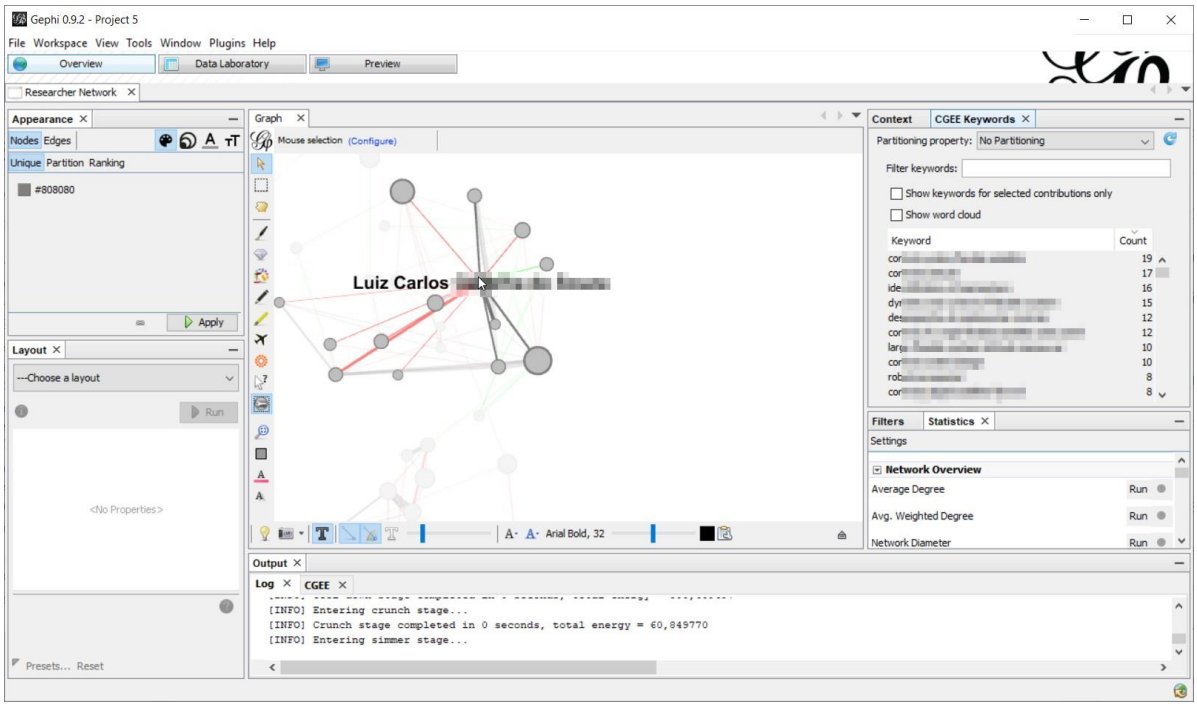


Figura 7.18: Exibição das palavras-chave do nó selecionado

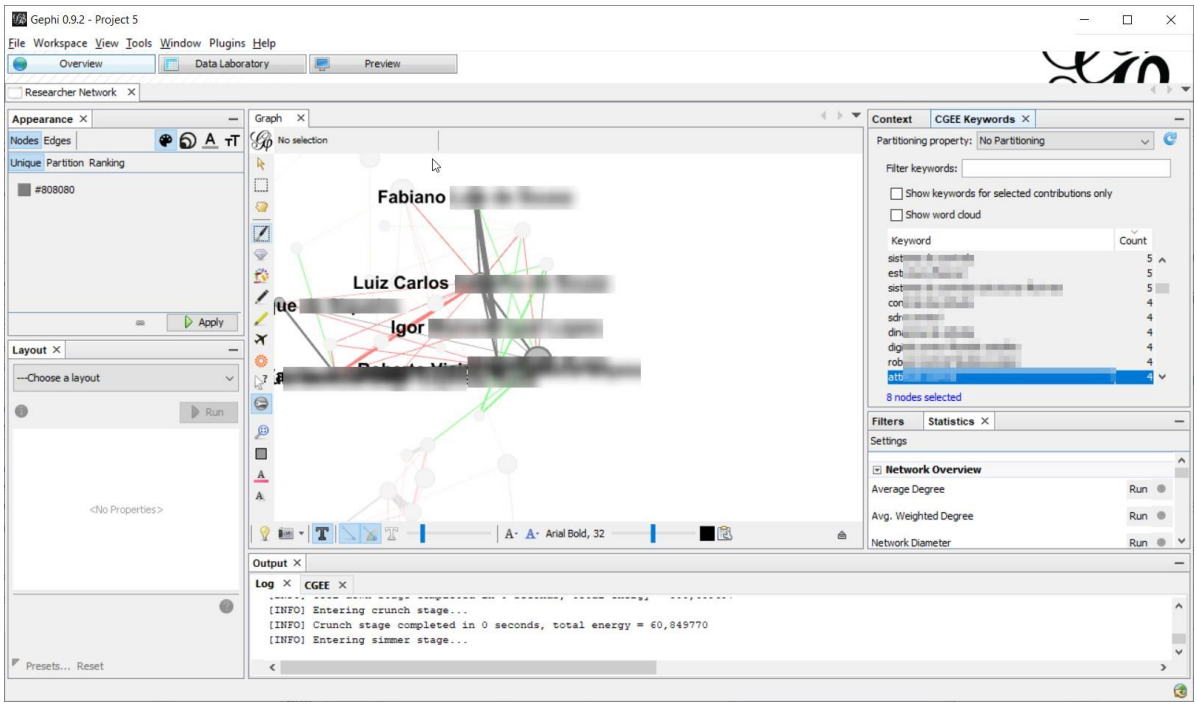


Figura 7.19: Exibição dos nós que usam a palavra-chave selecionada

A lista permite a seleção de uma única palavra-chave com um clique do botão esquerdo do mouse. Segu- rando o botão esquerdo, várias palavras-chave podem ser selecionadas, passando o mouse em cima dos nós. A mesma funcionalidade é obtida com um clique na primeira e na última palavra-chave, segurando a tecla Shift. Finalmente, várias palavras-chave podem selecionadas e desselecionadas independente- mente uma da outra, clicando nelas e segurando a tecla Ctrl.

7.4.4 Funcionalidades adicionais da lista de palavras-chave

Um clique com botão direito na lista de palavras-chave mostra a seguinte lista de funcionalidades adici- onais:

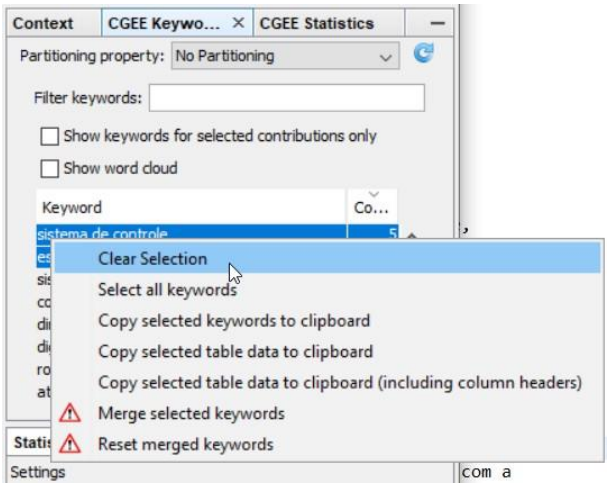


Figura 7.20: Funcionalidades adicionais na lista de palavras-chave

Os primeiros dois itens (“*Clear selection*” e “*Select all keywords*”) permitem retirar seleções prévias ou selecionar todos os itens da lista, considerando a funcionalidade descrita na seção anterior.

O terceiro item “*Copy selected keywords to clipboard*” copia as palavras-chave selecionadas para a área de transferência do sistema operacional, da qual elas podem ser coladas em programas de edição de textos e tabelas. O quarto item “*Copy selected table data to clipboard*” acrescenta as frequências e os outros valores numéricos exibidos. O próximo item “*Copy selected table data to clipboard (including column headers)*”, além da funcionalidade anterior, grava uma primeira linha com os nomes das colunas.

O item “*Merge selected keywords*” aparece apenas se mais que uma palavra-chave for selecionada na liste. Este item permite a junção de várias palavras-chave sinônimas. Selecionando essas palavras-chave e clicando em “*Merge selected keywords*”, o seguinte diálogo aparece na tela:

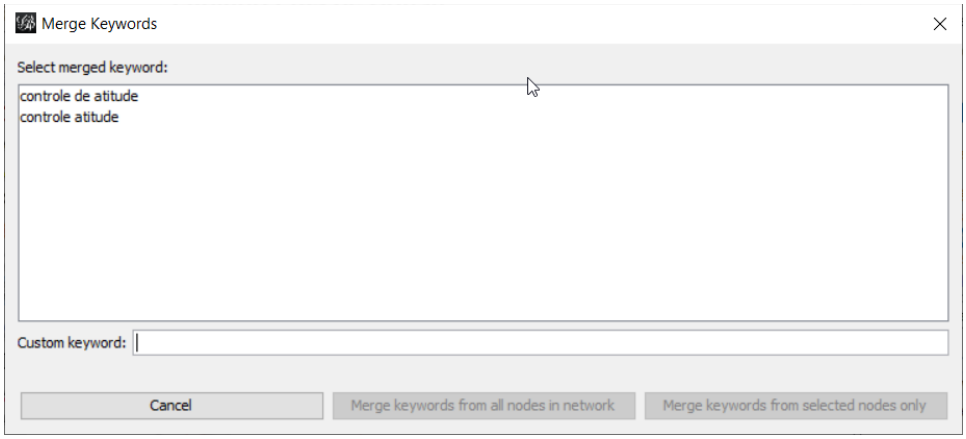


Figura 7.21: Diálogo de junção de palavras-chave

O usuário precisa selecionar qual das palavras-chave será a palavra-chave juntada. Todas as outras palavras-chave serão eliminadas:

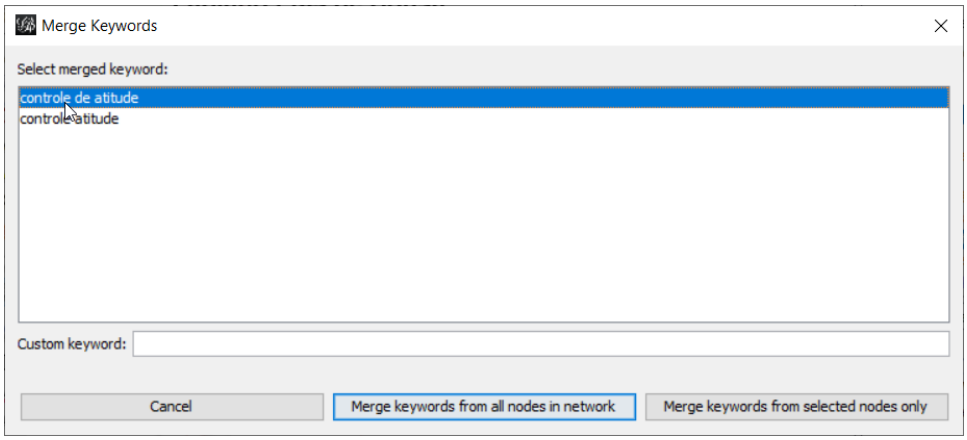


Figura 7.22: Seleção da palavra-chave juntada

Alternativamente, é possível digitar uma palavra-chave nova que substitui todas as palavras-chave sele- cionadas:

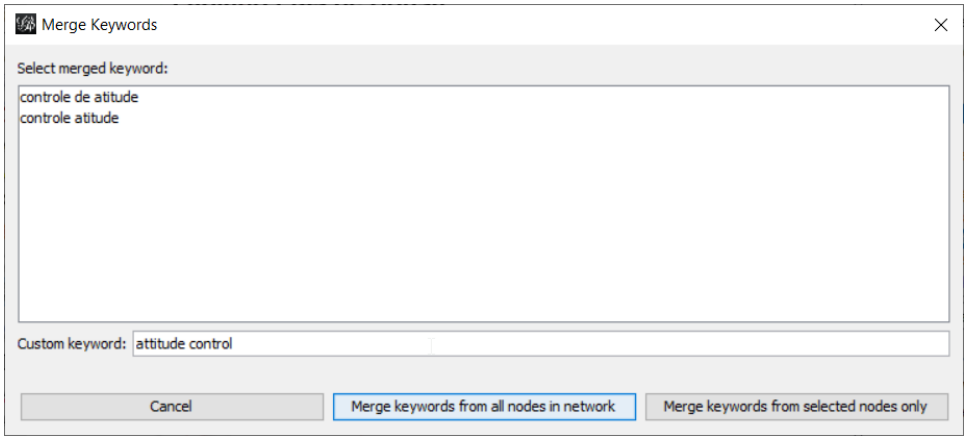


Figura 7.23: Especificação de uma palavra-chave nova

Finalmente, para realizar a junção das palavras-chave, o usuário precisa determinar se serão juntadas apenas as as palavras-chave que ocorrem nos nós selecionados ("*Merge keywords from selected nodes only*") ou se a operação deve ser realizada em todos os nós da rede ("*Merge keywords from all nodes in network*").

O item "*Reset merged keywords*" desfaz **todas** as operações de palavras-chaves juntadas e repõe a lista no estado original.

7.4.5 Nuvem de palavras-chave

A nuvem de palavras-chave exibe o conjunto de palavras-chave em uma única visualização em que o tamanho da palavra corresponde à frequência, à relevância ou à porcentagem de nós que contêm essa palavra.

Para exibir a nuvem de palavras, o usuário deve selecionar o item "*Show Word Cloud*" na janela de palavras-chave. Neste caso, aparece uma nova aba que mostra a nuvem de palavras:

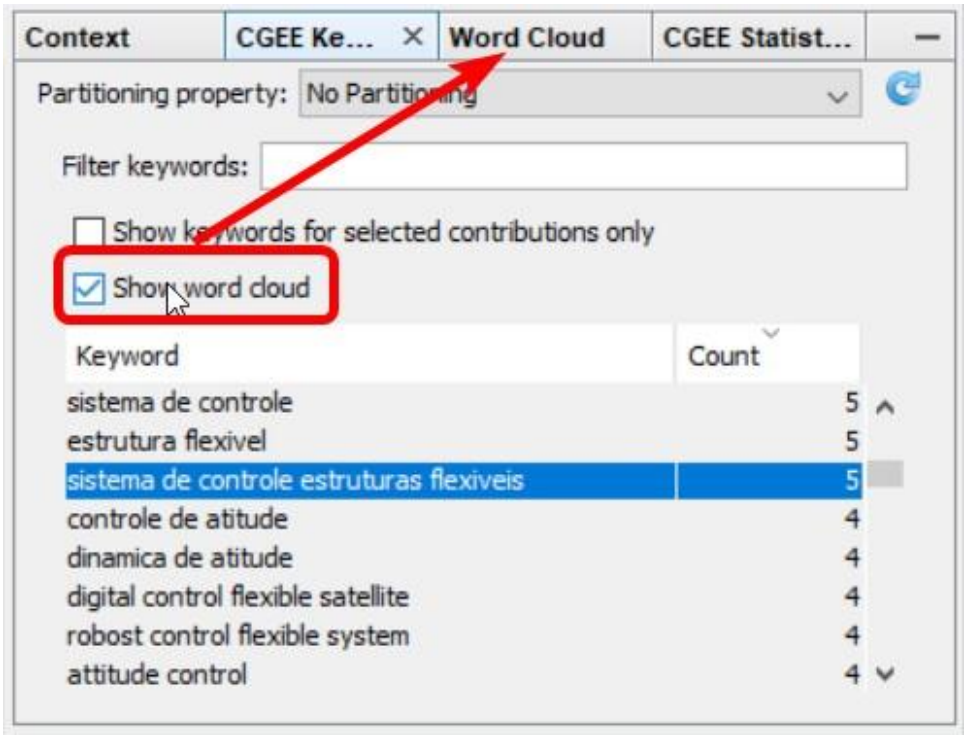


Figura 7.24: Habilitação da nuvem de palavras-chave

Clicando na aba "*Word Cloud*", a nuvem de palavras é exibida, junto com eventuais avisos sobre a falta de precisão:

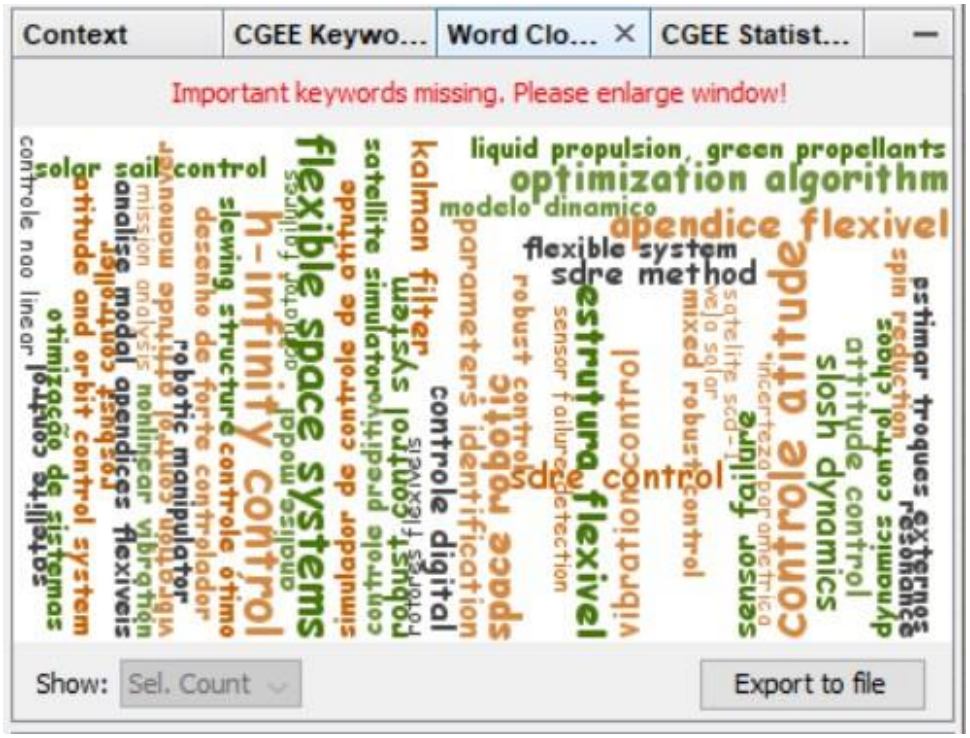
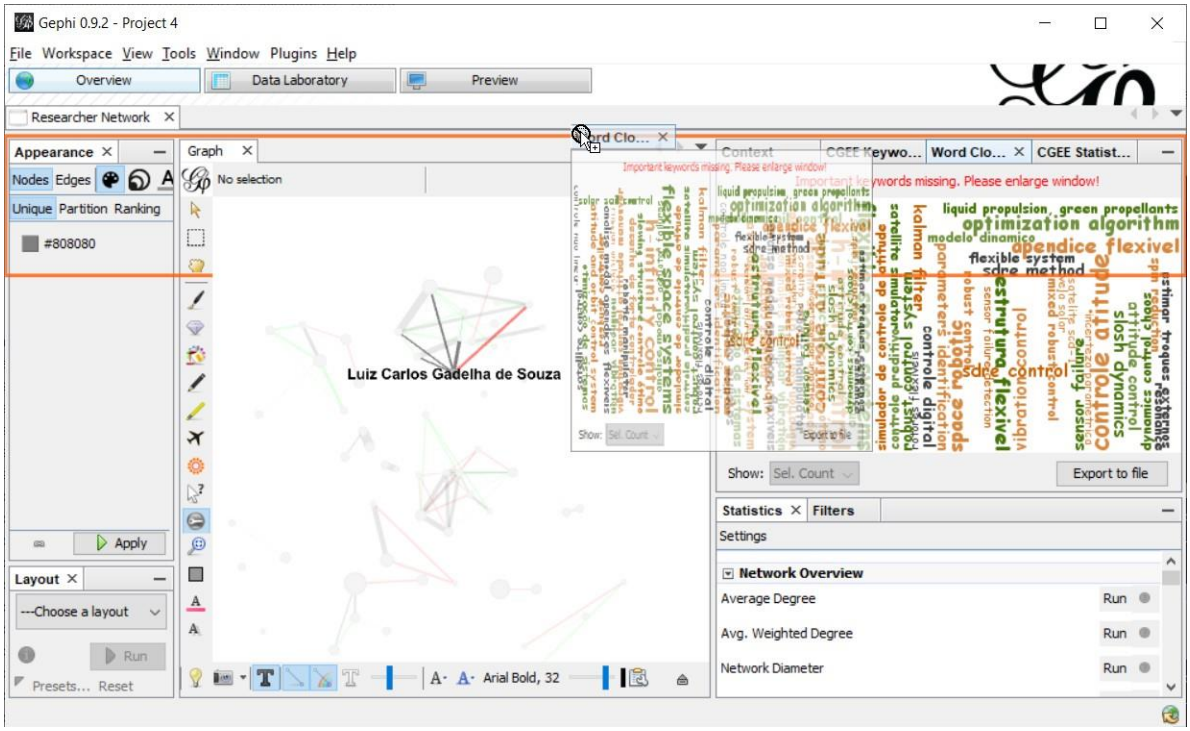


Figura 7.25: Nuvem de palavras-chave, com aviso de falta de precisão

Na criação da nuvem de palavras-chave deve ser observado que nem sempre é possível encaixar todas as palavras-chave que são importantes no espaço disponível. Caso isso ocorra, a janela deve ser aumentada. Clicando na aba "*Word Cloud*", é possível arrastar a janela para fora do aplicativo e tratá-la em separado dos outros elementos visuais:



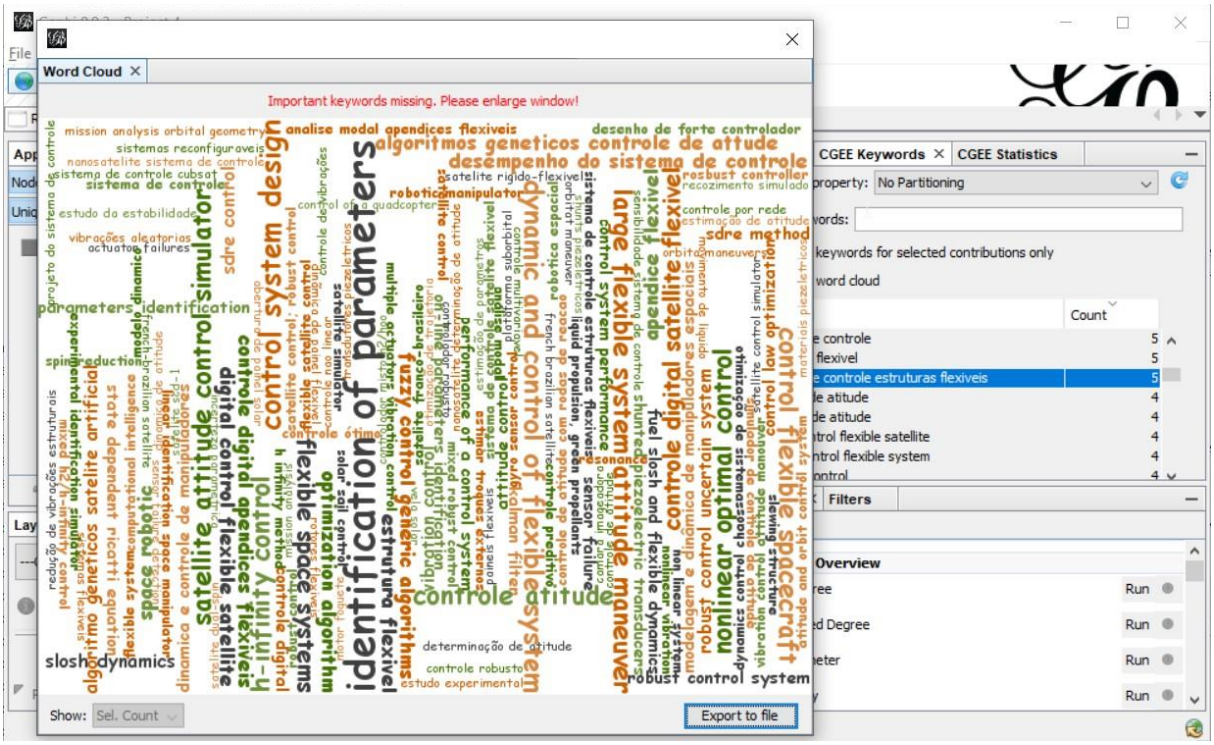


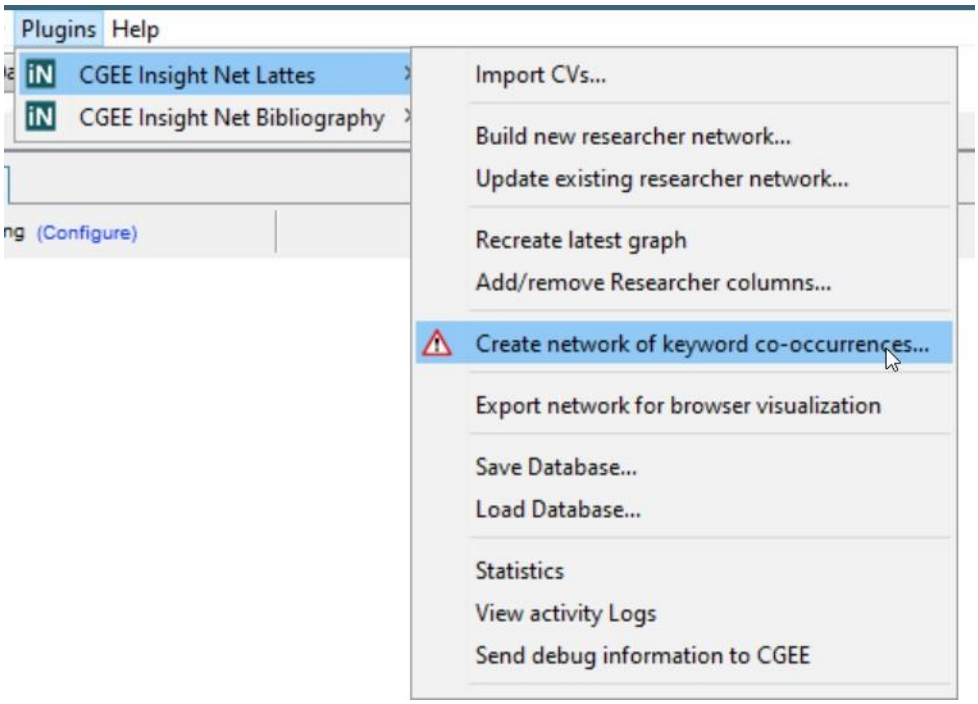
Figura 7.26: Separação da janela de nuvem de palavras

Se o espaço disponível na janela não permitir a exibição de, no mínimo, uma das 50 palavras-chave mais importantes de acordo com o critério selecionado, é exibida a mensagem "*Important keywords missing. Please enlarge window.*" Caso contrário, aparece a mensagem "*Complete. xx% of relevant keywords shown*". É importante notar que a nuvem sempre é montada na sequência decrescente dos valores do critério selecionado (frequência, relevância ou porcentagem de nós). Entretanto, palavraschave de maior valor usam mais espaço e podem não mais caber na área disponível da nuvem, enquanto palavras de menor valor ocupam menos espaço e assim podem ser incluídas.

7.5 Criação de redes de co-ocorrências de palavras-chave

Como funcionalidade adicional, o *CGEE Insight Net* permite a criação de redes de co-ocorrências de palavras-chave. Isso significa que existem arestas entre palavras-chaves que ocorrem juntos em um ou mais contribuições bibliográficas[1]. Cada contribuição em que as ambas as palavras-chave ocorrem aumenta o peso da aresta entre as duas palavras-chave em um.

As redes de palavras-chave podem ser criadas por qualquer tipo de rede préviamente criada no Gephi. A opção "*Create network of keyword co-occurrences*" abre o seguinte diálogo:



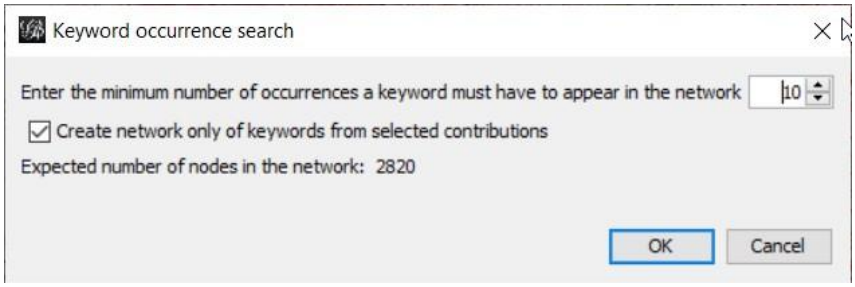


Figura 7.27: Funcionalidade para criar redes de palavras-chave

Como a quantidade de palavras-chave costuma ser alta, apenas as palavras-chaves com o maior número de ocorrências podem ser consideradas. Essa quantidade mínima de ocorrências de cada palavra-chave pode ser especificada.

A segunda opção refere-se ao escopo de extração de palavras-chave. Podem ser consideradas as palavras-chaves de **todas** as contribuições bibliográficas ou apenas aquelas que foram consideradas na criação da rede anterior (de pesquisadores ou de referências bibliográficas).

[1] *Contribuições bibliográficas* são as referências bibliográficas importadas pelos módulos “BibTeX” e “Bibliografia gené- rica”, bem como os artigos, capítulos de livros e trabalhos em eventos que constam nos Currículos Lattes.

7.6 Eliminação interativa de nós da rede e do banco de dados

Conforme descrito na seção Seção 4, a rede de pesquisadores ou de referências bibliográficas é criada a partir do conteúdo de banco de dados. Isso significa que a eliminação de um nó com as funcionalidades do Gephi (clicando no nó com botão direito e selecionando "*Delete*"), o exclui apenas do grafo, mas **não** do banco de dados. Se a rede for calculada novamente ou recuperada a partir da função "*Recreate latest graph* (ver Seção 8.1), o nó reaparece.

O *CGEE Insight Net* possui duas funcionalidades que permitem a exclusão interativa de nós do banco de dados, que podem ser executadas com um clique do botão direito no nó dentro da visualização do grafo ou no laboratório de dados:

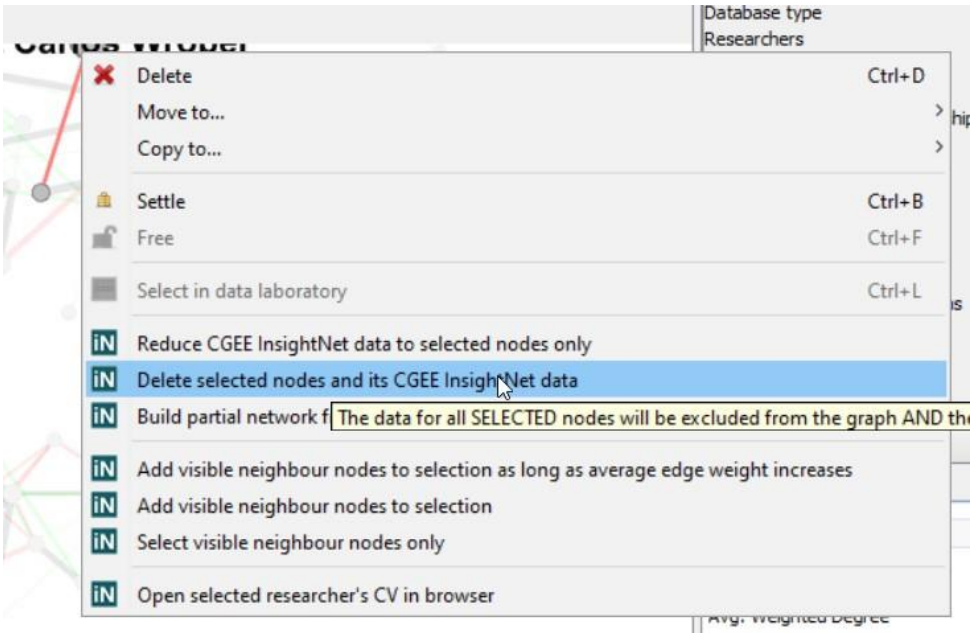


Figura 7.28: Eliminação interativa de nós do banco de dados na visualização do grafo

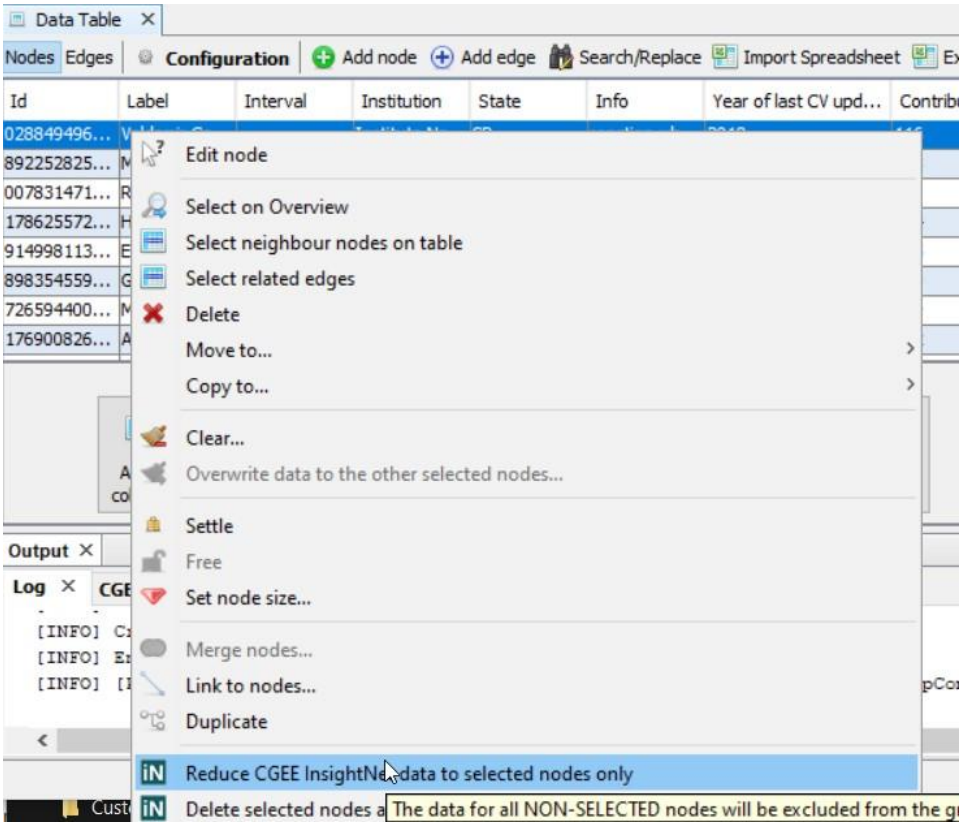


Figura 7.29: Eliminação interativa de nós do banco de dados no laboratório de dados

As duas funções "*Reduce CGEE InsightNet data to selected Nodes only*" e "*Delete selected nodes and its CGEE InsightNet data*" possuem objetivos complementares. Ambas requerem um conjunto de nós selecionados. Com um clique na função "*Reduce CGEE InsightNet data to selected Nodes only*", apenas esses nós selecionados permanecem no banco de dados, todos os nós não selecionados serão eliminados do grafo e do banco de dados.

Já um clique na função "*Delete selected nodes and its CGEE InsightNet data*" elimina apenas os nós selecionados e os não selecionados permanecem no grafo e no banco de dados.

> A eliminação de um ou mais nós da rede altera os pesos das arestas entre os nós e a rede de similaridade semântica precisa ser recalculada. Por esse motivo, todas as arestas da rede são eliminadas e a rede precisa ser recalculada.

7.7 Criação de uma nova rede a partir do subconjunto de nós se- lecionados

A funcionalidade "*Build partial network from selected nodes*", também disponível com clique com botão direito após a seleção de vários nós na visualização do grafo ou no laboratório de dados, permite a criação de uma nova rede em que constam apenas os nós selecionados.

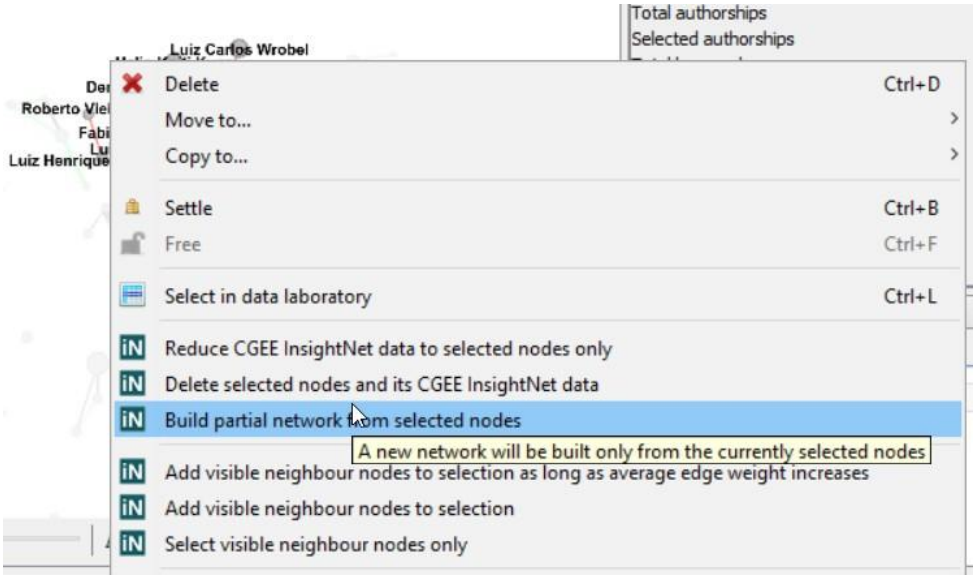


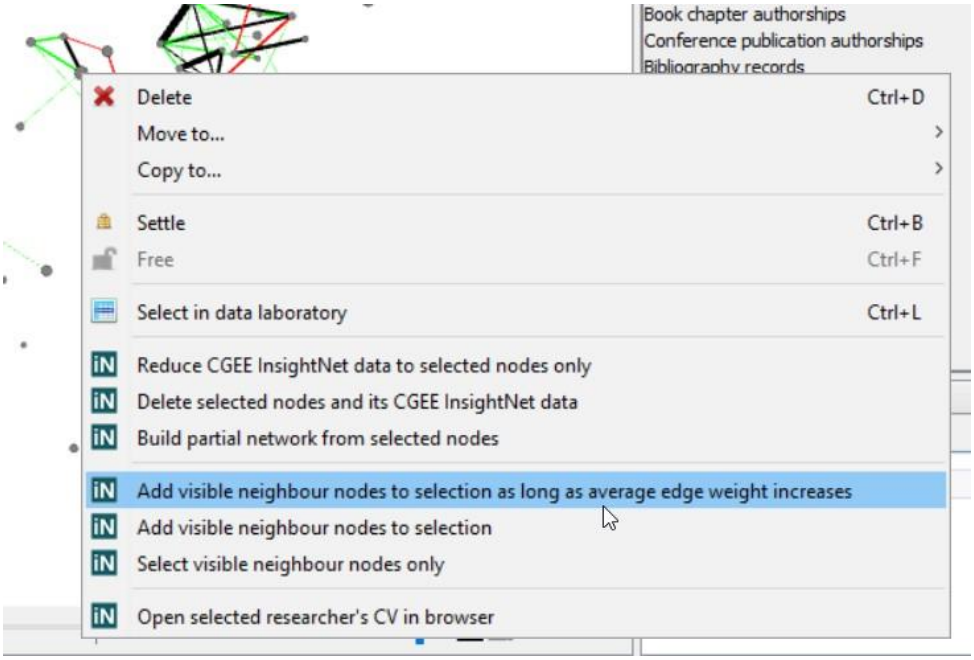
Figura 7.30: Criação de uma rede a partir do subconjunto de nós selecionados

Os nós que não foram selecionados vão permanecer no banco de dados, mas não farão parte da rede criada.

Esta funcionalidade permite a análise de várias sub-redes sem precisar realizar novas importações de dados.

7.8 Seleção interativa de nós vizinhos na rede

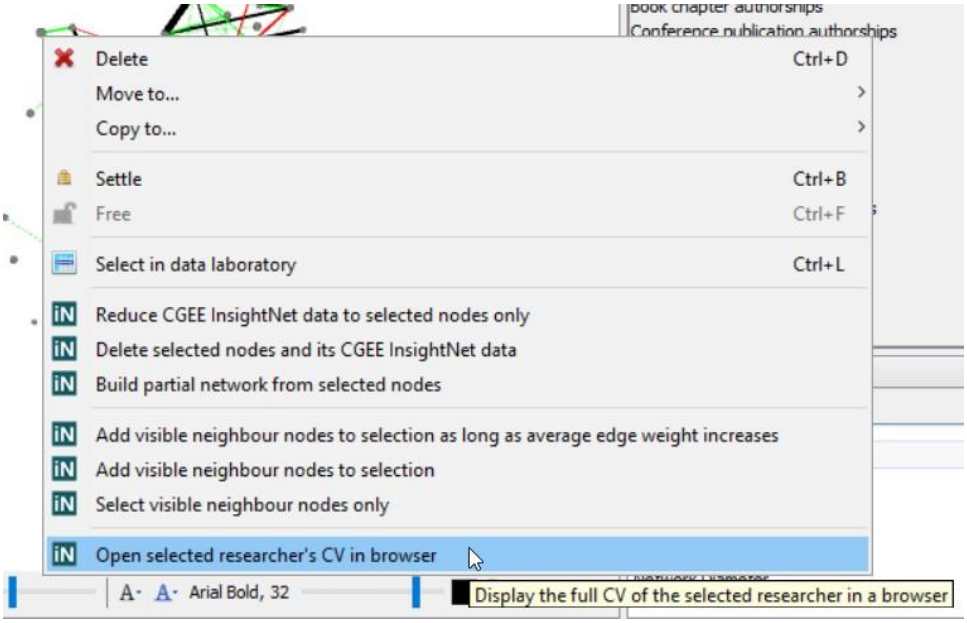
A partir de um ou mais nós selecionados, o *CGEE Insight Net* permite a seleção dos nós vizinhos com um clique do botão direito do mouse no nó dentro da visualização do grafo ou no laboratório de dados. A função "*Add visible neighbour nodes to selection*" acrescenta ao conjunto de nós selecionados todos os vizinhos visíveis desses nós. Já a função "*Select visible neighbour nodes only*" substitui o conjunto de nós selecionado pelos nós vizinhos.

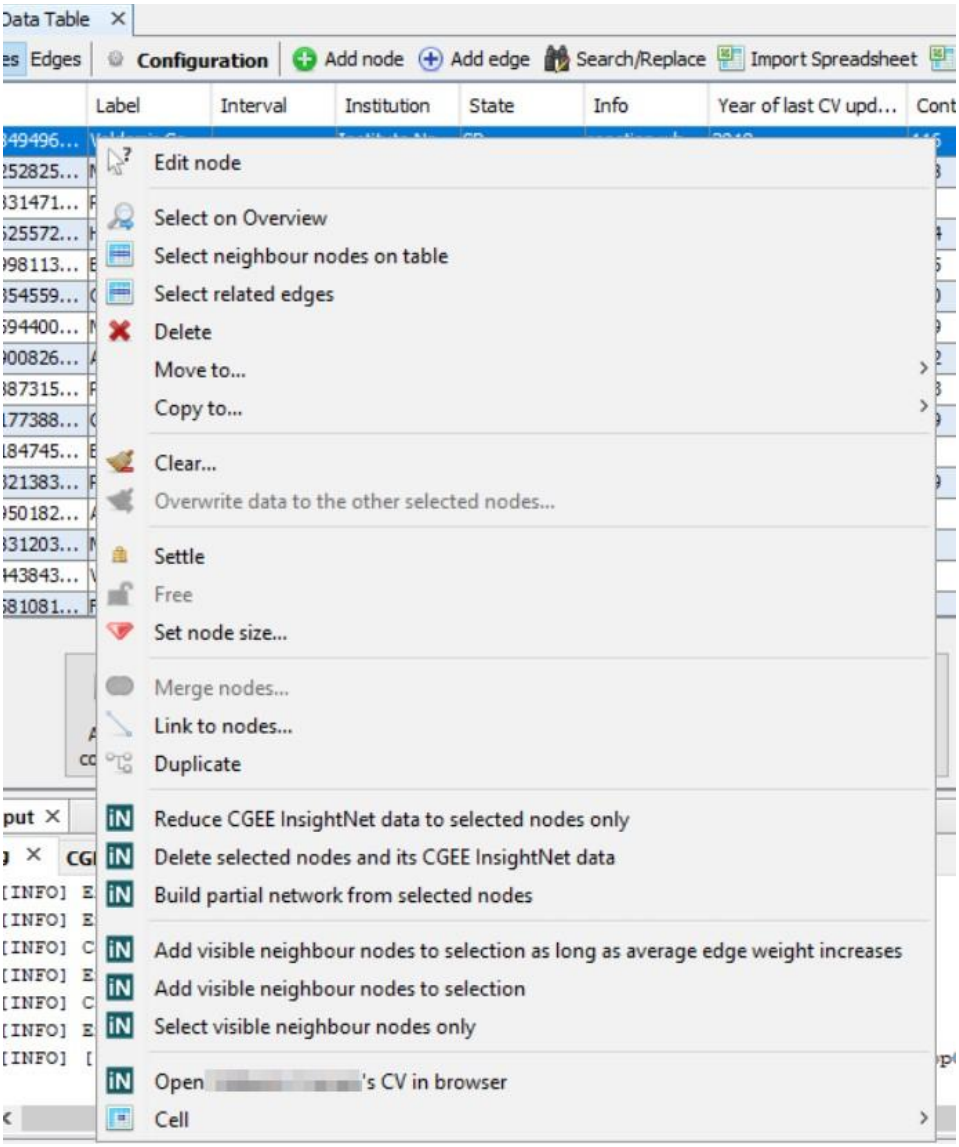


Já a funcionalidade "*Add visible neighbour nodes to selection as long as average edge weight increases*" repete o processo de adicionar vizinhos enquanto o peso média das arestas não diminua. Desta forma, o processo termina quando os nós adicionados não agregam mais informações relevantes ao subconjunto de nós selecionados.

7.9 Visualização interativa do currículo de pesquisadores no brow- ser

Em redes de pesquisadores, o Currículo Lattes de um pesquisador pode ser aberto no browser, clicando com o botão direito do mouse no nó da rede e selecionando a funcionalidade "*Open selected researcher's CV in browser*" na visualização do grafo ou "*Open <name>'s CV in browser*" no laboratório de dados. Essa funcionalidade abre o site do CNPq e, portanto, depende do acesso à internet.




Data Table
Edges
Configuration
Add node
Add edge
Search/Replace
Import Spreadsheet
Label
Interval
Institution
State
Info
Year of last CV upd...
Edit node
Select on Overview
Select neighbour nodes on table
Select related edges
Delete
Move to...
Copy to...
Clear...
Overwrite data to the other selected nodes...
Settle
Free
Set node size...
Merge nodes...
Link to nodes...
Duplicate
Reduce CGEE InsightNet data to selected nodes only
Delete selected nodes and its CGEE InsightNet data
Build partial network from selected nodes
Add visible neighbour nodes to selection as long as average edge weight increases
Add visible neighbour nodes to selection
Select visible neighbour nodes only
Open [redacted]'s CV in browser
Cell

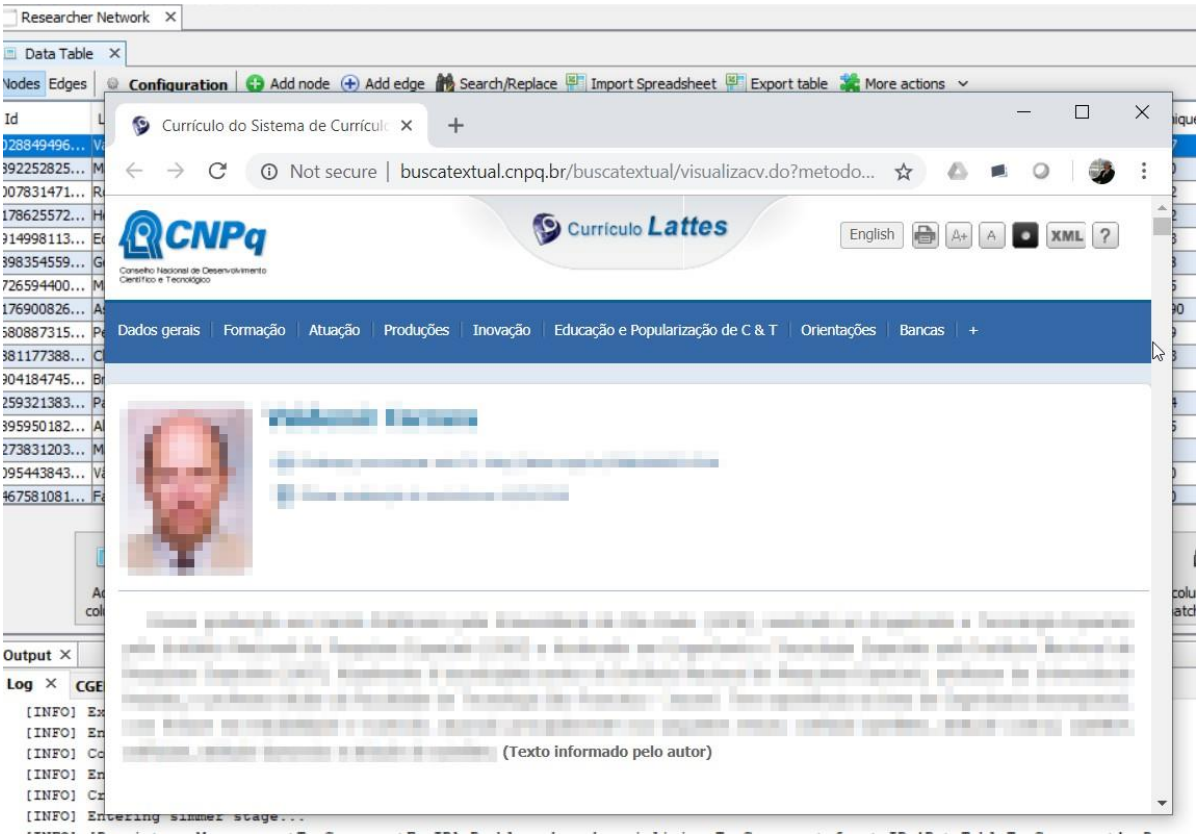
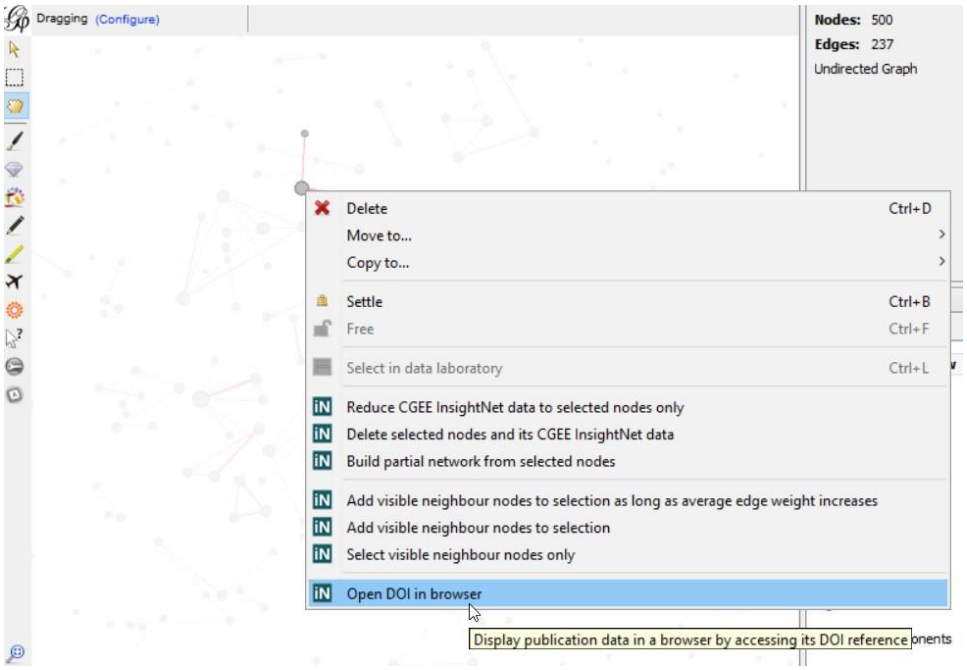
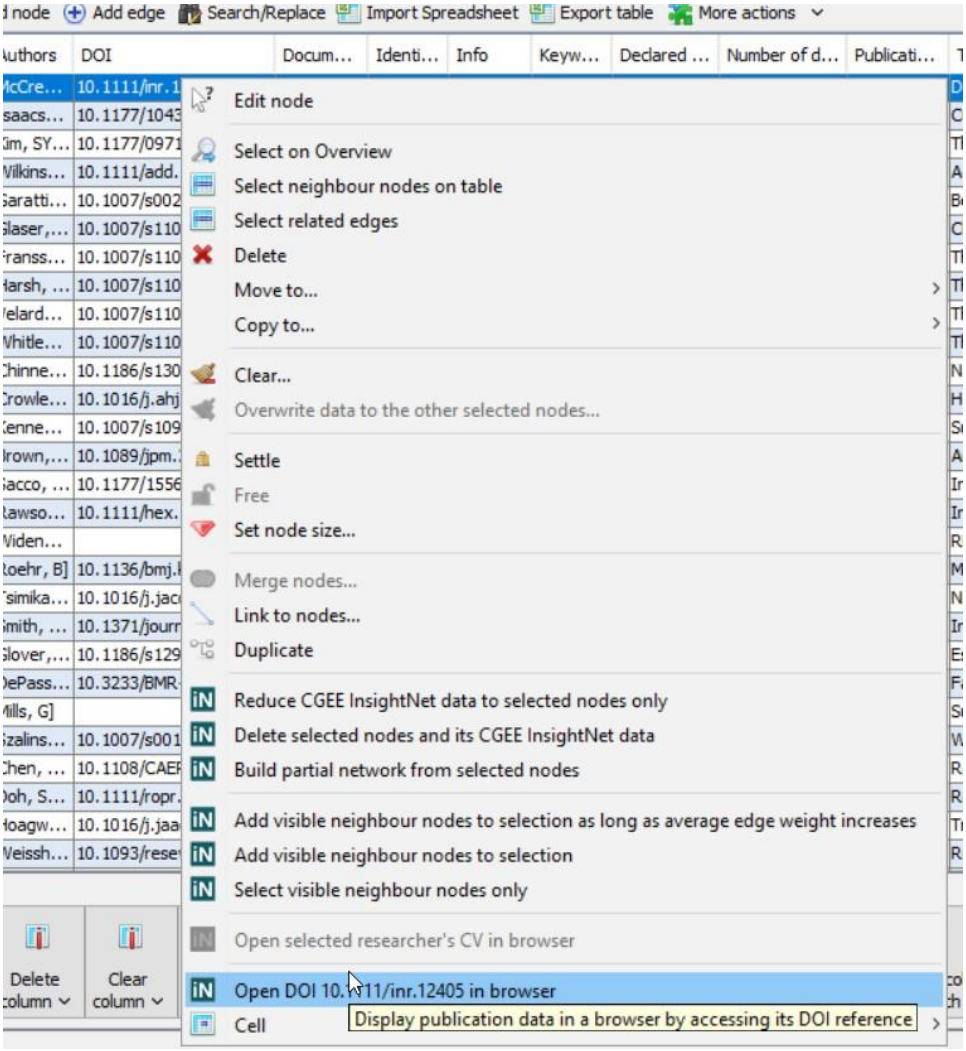


Figura 7.31: Visualização do currículo do pesquisador no Browser

7.10 Visualização interativa de contribuições bibliográficas por DOI no browser

Em redes de contribuições bibliográficas, o *Document Object Identifier* (DOI) permite acesso direto à referência bibliográfica e, dependendo da editora, também ao conteúdo. Um clique com botão direito do mouse na publicação no grafo ou no laboratório de dados mostrará o menu de *popup* com a opção "*Open DOI <doi> in browser*". Clicando nessa opção, a página do DOI é exibido no browser do usuário.




Add edge
Search/Replace
Import Spreadsheet
Export table
More actions
Authors
DOI
Docum...
Identi...
Info
Keyw...
Declared ...
Number of d...
Publicati...
Edit node
Select on Overview
Select neighbour nodes on table
Select related edges
Delete
Move to...
Copy to...
Clear...
Overwrite data to the other selected nodes...
Settle
Free
Set node size...
Merge nodes...
Link to nodes...
Duplicate
Reduce CGEE InsightNet data to selected nodes only
Delete selected nodes and its CGEE InsightNet data
Build partial network from selected nodes
Add visible neighbour nodes to selection as long as average edge weight increases
Add visible neighbour nodes to selection
Select visible neighbour nodes only
Open selected researcher's CV in browser
Open DOI 10.1111/inr.12405 in browser
Display publication data in a browser by accessing its DOI reference
Cell
Delete column
Clear column

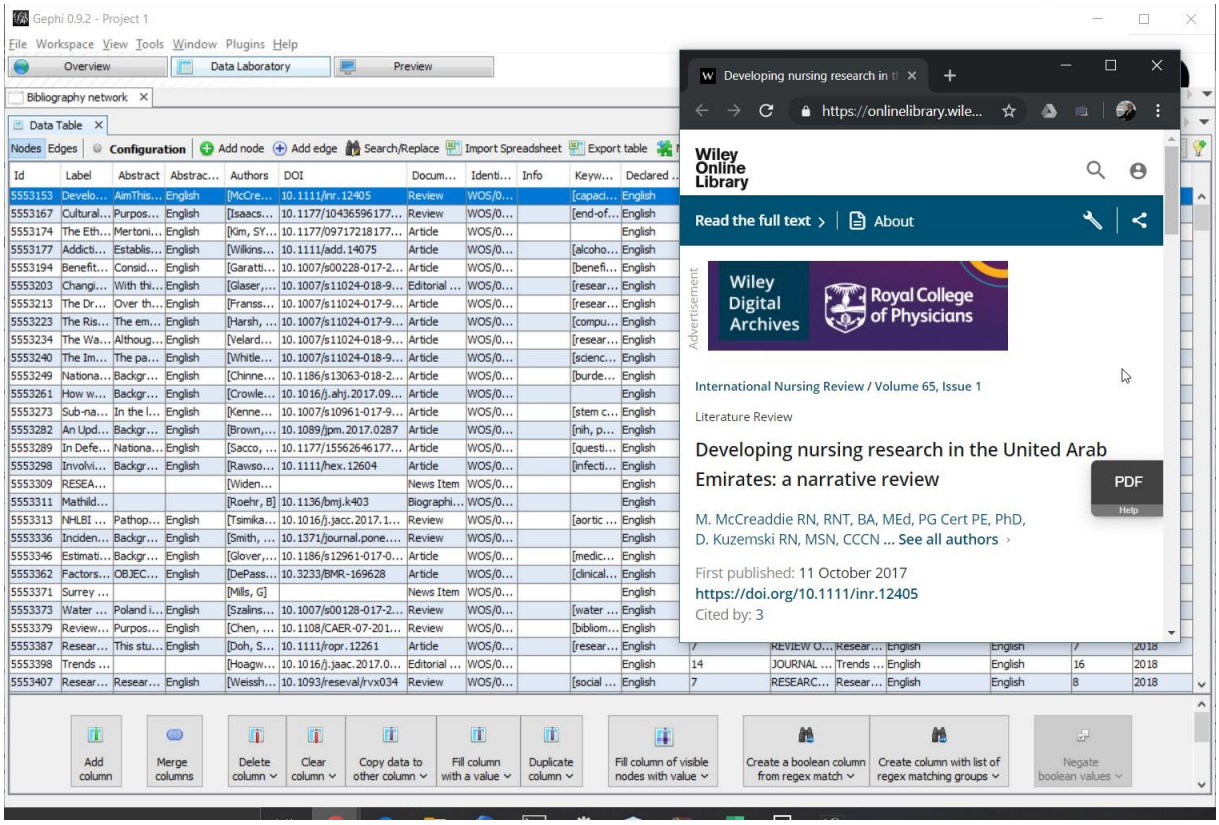


Figura 7.32: Visualização da referência bibliográfica por DOI no Browser

CAPÍTULO 8

Funcionalidades comuns de apoio

As funcionalidades descritas nessa seção se aplicam a todos os tipos de rede e são replicadas em todos os sub-menus do CGEE Insight Net. Entretango, essa seção se limita à demonstração dos diagramas dos itens apenas no sub-menu "*CGEE Insight Net Lattes*".

8.1 Recuperação do grafo a partir das informações que constam no banco de dados

A seleção das contribuições e a pesquisa de similaridade são realizadas exclusivamente no banco de dados, de acordo com a tabela na Seção 4. O último passo (passo 4 da tabela) gera o grafo a partir das informações que constam no banco de dados.

Esse passo pode ser executado isoladamente e permite a recuperação do grafo sem precisar executar a demorada pesquisa por similaridade, uma vez que a ferramenta Gephi não permite desfazer algumas ações realizadas. O item *Plugins* > *CGEE Insight Net . . .* > *Recreate latest graph* extrai as informações consolidadas do banco de dados e cria um novo grafo no Gephi:

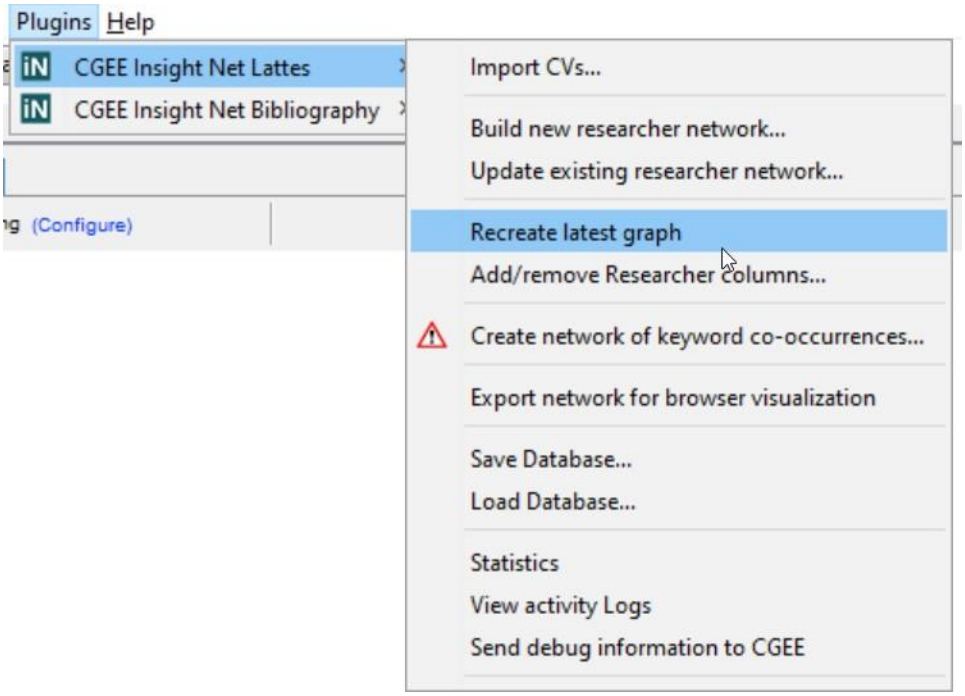


Figura 8.1: Recuperação do grafo

Selecionando essa opção, o grafo é criado a partir das informações no banco de dados, da mesma forma em que o passo 4 da tabela da Seção 4 é executado:

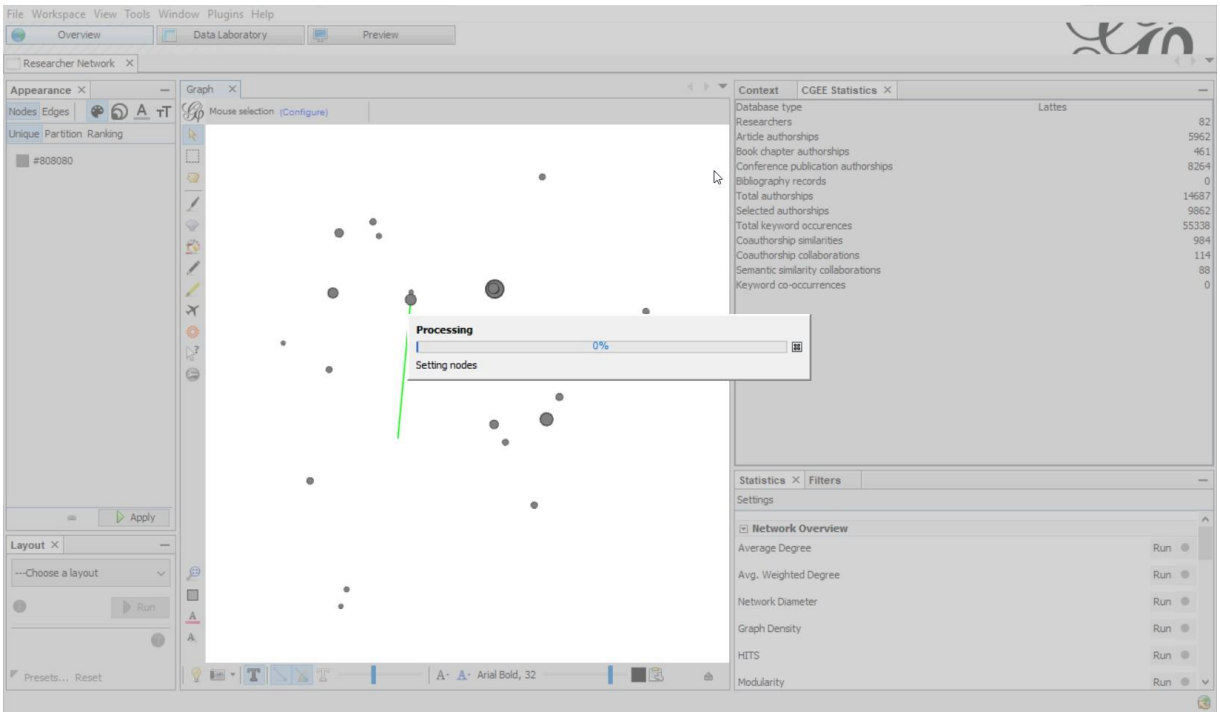


Figura 8.2: Recuperação do grafo

8.1. Recuperação do grafo a partir das informações que constam no banco de dado1s09

8.2 Cópia e recuperação do banco de dados

Conforme mencionado na Seção 4, todos os dados relevantes selecionados para análise dos currículos constam no banco de dados e apenas no final do processamento são transformados em um grafo visível. Assim, o banco de dados é o repositório principal das informações – o grafo gravado no arquivo `.gephi` é apenas uma representação visual das informações computadas.

O *CGEE Insight Net* permite a gravação do banco de dados em um arquivo do tipo `.cge` e a respectiva recuperação a partir dos itens "*Save Database*" e "*Load Database*" do menu *Plugins* > *CGEE Insight Net . . .* :

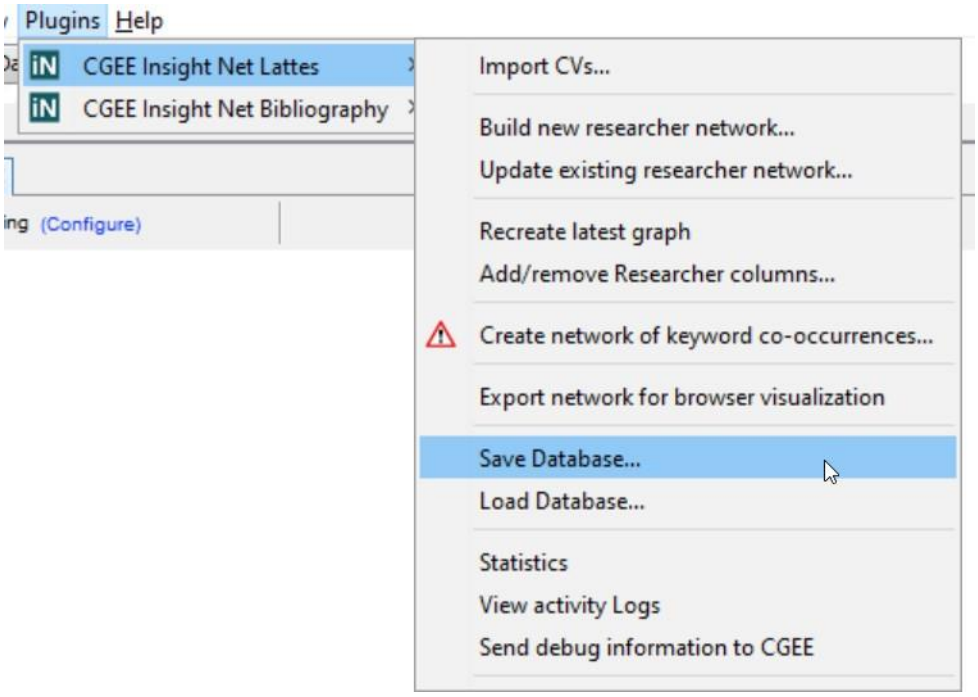


Figura 8.3: Funcionalidades de gravação e recuperação do banco de dados

Selecionando o item "Save Database", o sistema apresenta um diálogo e solicita a definição do arquivo a ser gravado:

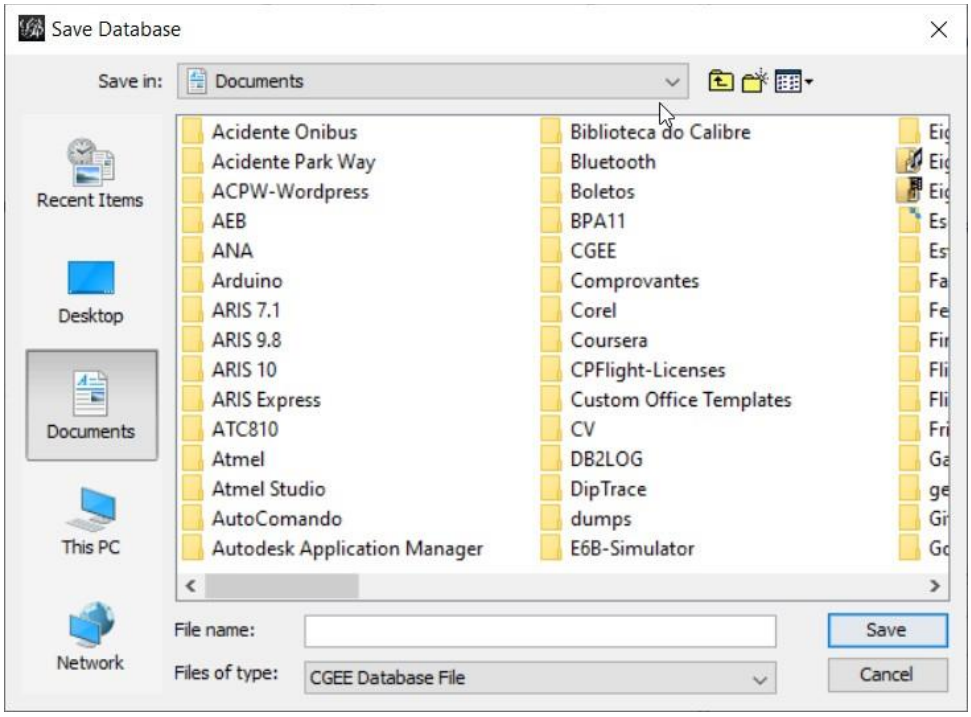


Figura 8.4: Especificação do arquivo de backup do banco de dados

Se o usuário especificar um arquivo válido e clicar em "*Save*", o backup do banco de dados é gerado:

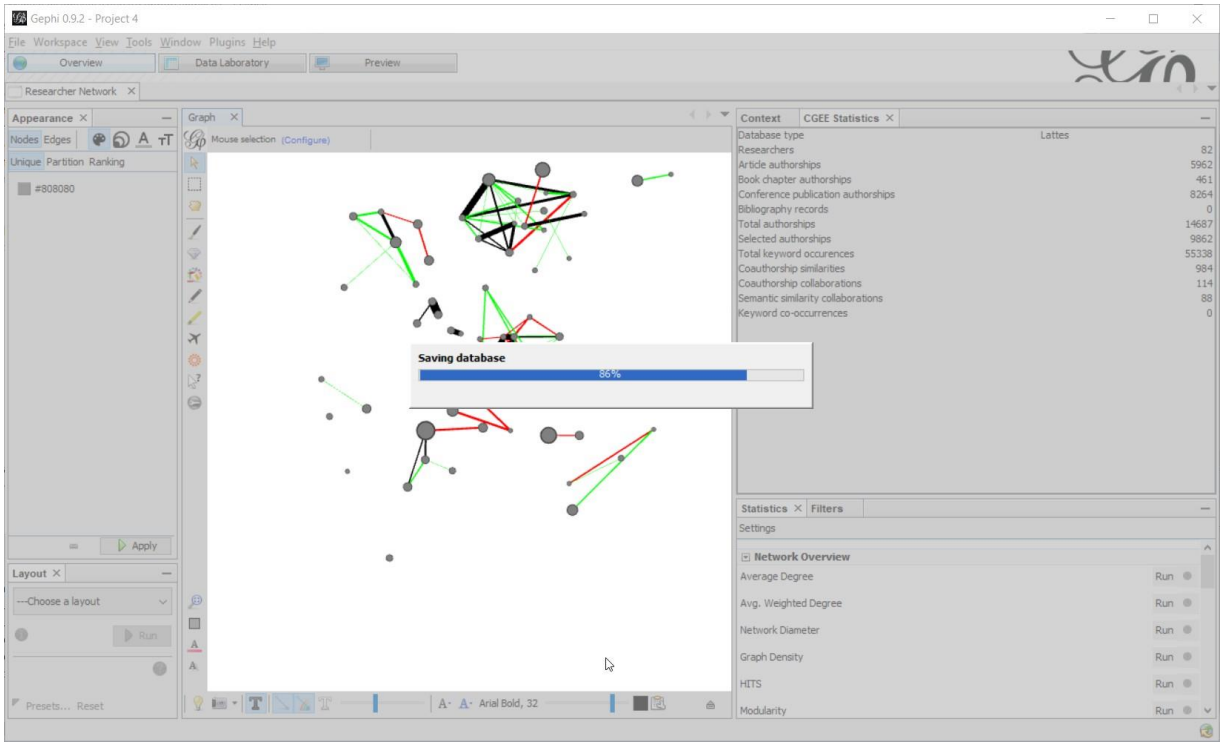
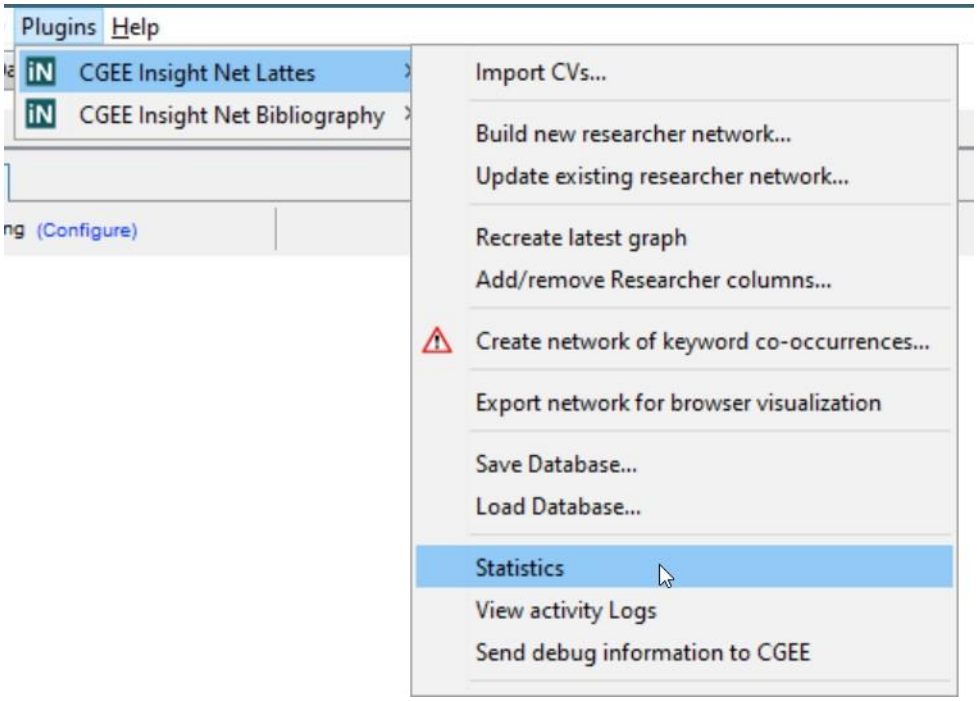


Figura 8.5: Geração do backup da base de dados

8.3 Estatísticas do banco de dados

O *CGEE Insight Net* permite exibir uma estatística do banco de dados com a opção *Plugins* > *CGEE Insight Net . . .* > *Statistics*, que abre a janela de estatística do banco de dados:



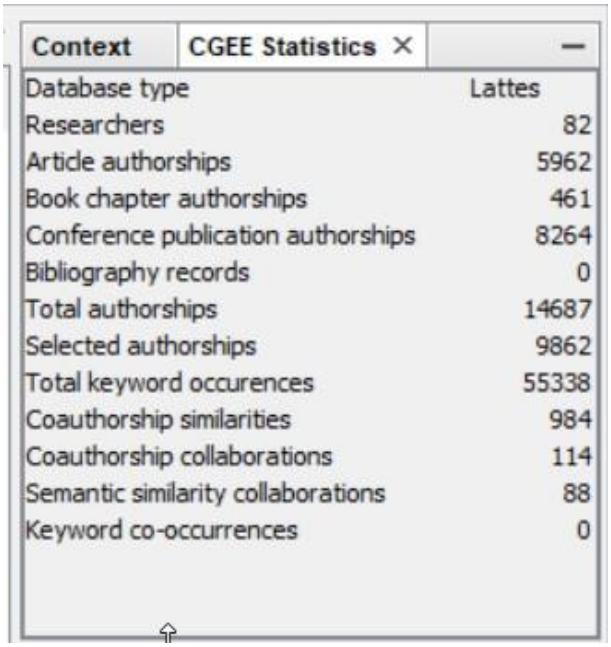


Figura 8.6: Janela de estatística

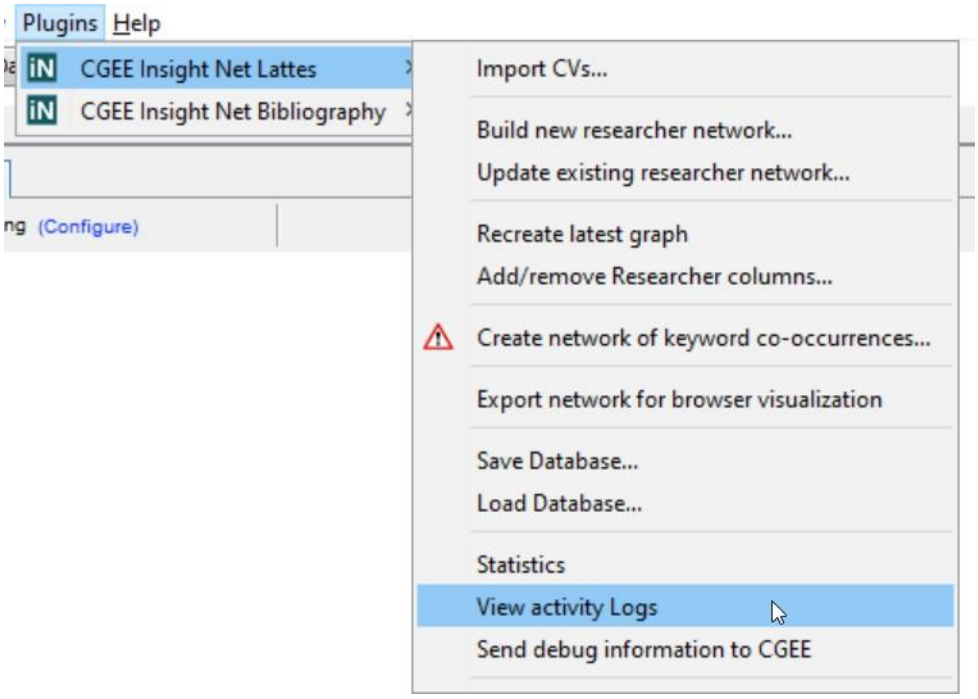
Essa janela mostra o tipo e a quantidade de registros para vários tipos de dados no banco. Percebe- se que depois da importação e antes da geração da rede, os valores para "*Selected*

Contributions", "*Coauthorship Similarities*", "*Coauthorship collaborations*", "*Semantic similarity collaborations*" e

"*Keyword co-occurrences*" ficam com o valor zero, já que essas entidades são geradas apenas durante a formação da rede.

8.4 Protocolos de execução

Conforme descrito na Seção 3, o *CGEE Insight Net* gera diversos registros de protocolo. Eles podem ser visualizados com a opção *Plugins* > *CGEE Insight Net* > *View Logs*, que abre a janela de protocolo:



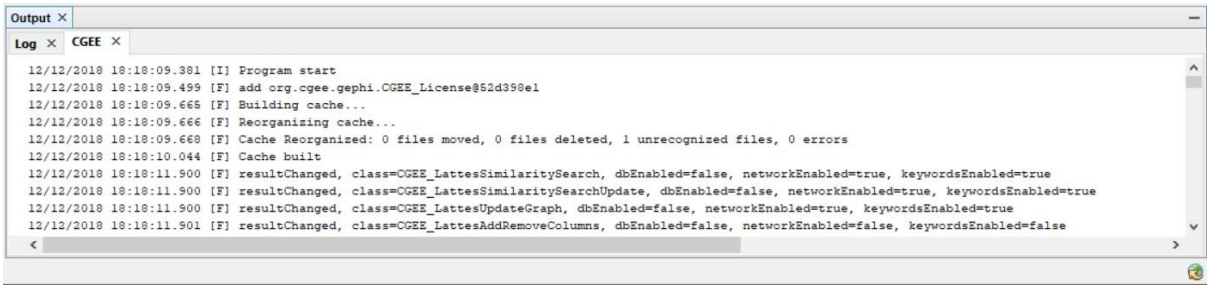


Figura 8.7: Protocolo de execução

A janela possui duas abas:

- Na aba "*Log*" são exibidas mensagens do proprio ambiente do Gephi, sem relação ao *CGEE Insight Net*
- Já a aba "CGEE" exibe os registros de protocolo de execução do *CGEE Insight Net*. Cada linha nesta aba "CGEE" é marcada com data e hora e com o tipo de registro:

Típo	Significado
[E]	Erro severo, integridade dos dados não garantida
[!]	Aviso importante
[I]	Informação sobre a execução do CGEE Insight Net
[F]	Informações detalhadas sobre a execução do CGEE Insight Net
[f]	Informação de depuração
[*]	Informação detalhada de depuração

8.5 Envio de protocolo de execução

O *CGEE Insight Net* permite enviar dados sobre a execução do *Gephi* e do *plugin* ao CGEE para facilitar a análise de possíveis problemas. A opção "*Send debug information to CGEE*" está disponível em todos os menus de *plugin*:

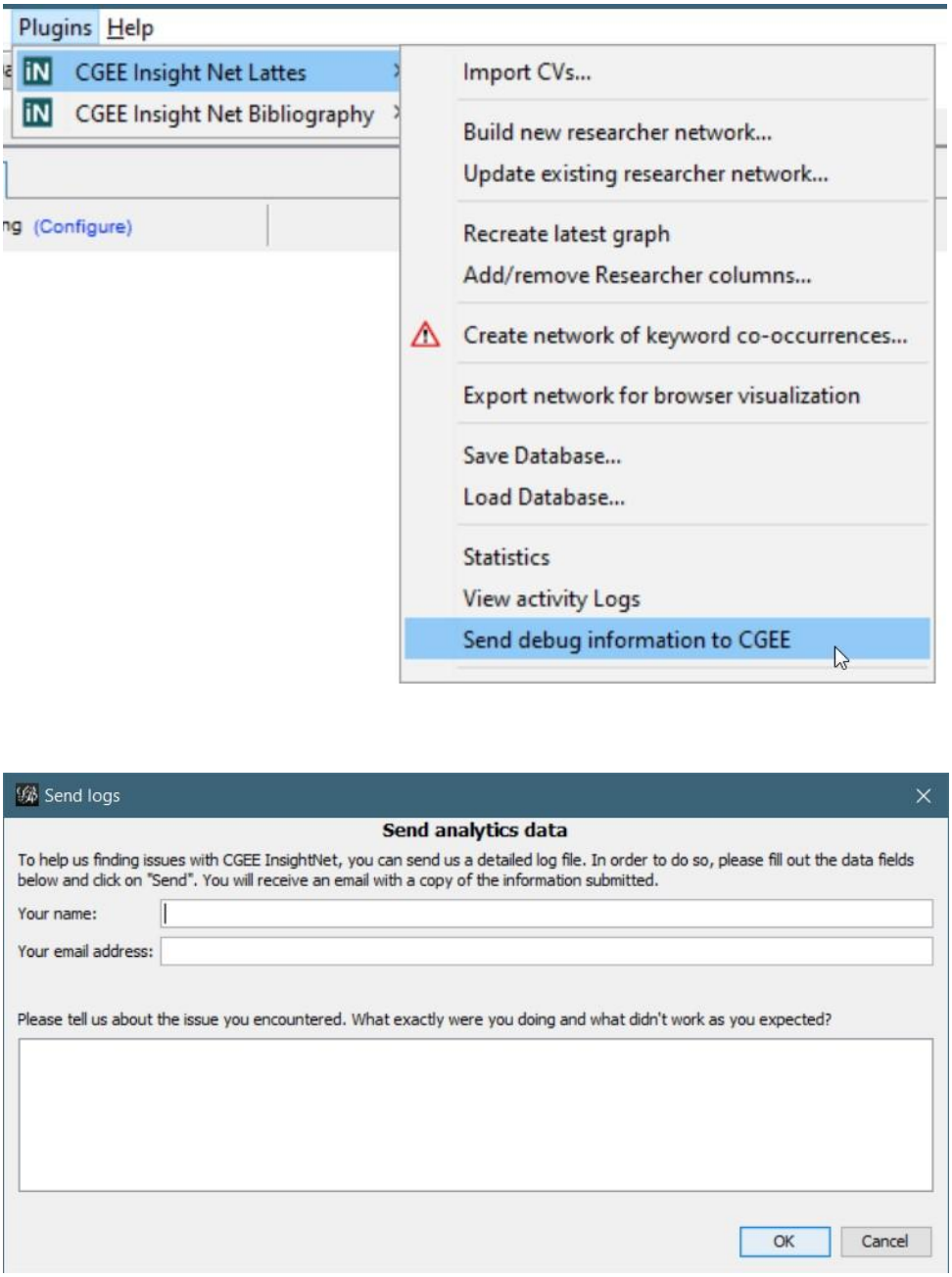


Figura 8.8: Enviar protocolos de execução

Caso a execução do *Gephi* for interrompida sem fechar o programa corretamente, o seguinte diálogo é exibido na próxima vez que o *Gephi* for iniciado, que oferece ao usuário enviar os protocolos da última execução:

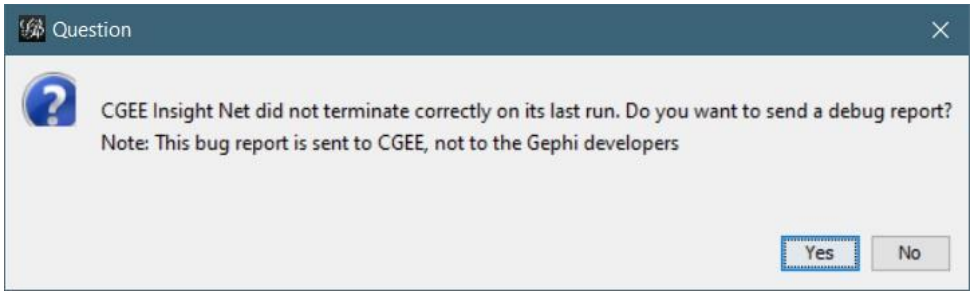


Figura 8.9: Diálogo que aparece depois da terminação forçada do *Gephi*

Referências Bibliográficas

[Assort] M. Newman, “*7.13 HOMOPHILY AND ASSORTATIVE MIXING*,” em *Networks. An Intro- duction*, New York, Oxford University Press, 2010.

[Percolation] A.-L. Barabási, “*8.2 Percolation Theory ff*”, em *Network Science*, Cambridge, Cambridge University Press, 2016, p. 273ff.