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

Exploração de dados e visualização de informação

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, 2020

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, 2020.

134p.: 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. Exploração de dados e visualização de informação. 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.

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
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 de interesse em 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, portanto os requisitos se concentram 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 necessários 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 Insight Net e seus resultados exportados para o InsightNet Browser. Com esta conexão entre suas três principais ferramentas de análise de textos, o Centro tem a possibilidade de realizar análises de redes e mineração de dados com documentos do acervo próprio (integrado ao InsightData), de currículos Lattes e de metadados de algumas das principais bases de dados de produção acadêmica e de patentes disponíveis. Como atividades centradas no

aprimoramento das ferramentas iN, iNB, iD e metodologias associadas, destacam-se:

a) Para o registro de origem de dados, exibição, na janela de importação, do número total de currículos disponíveis e a data da última atualização do espelho da base Lattes do CGEE.

b) Implementação da funcionalidade de importação de autores a partir de DOI (*digital object identifier*), o identificador único de artigos científicos. Esta funcionalidade permite que, dada uma seleção inicial de currículos, o usuário recupere todos os coautores de produções que tenham DOI.

c) Implementação de melhorias na exportação para programas gerenciadores de planilhas, como o Excel, de metadados de contribuições selecionadas a partir de currículos baixados. Essas melhorias viabilizaram a possibilidade de recuperação de coautores apenas de artigos selecionados a partir de filtros do próprio programa de gerenciamento de planilhas, como instituição do coautor, ano de publicação e, provavelmente o mais útil dos filtros, palavras-chave dos artigos, quando houver esse registro. Novos métodos de busca recursiva de currículos a partir de coautorias com bases anteriormente baixadas (*snowball sampling*) foram testados. O exemplo prático mais marcante foi o levantamento de instituições que tem coautores de artigos das unidades de pesquisa do MCTI, que pode evoluir para a elaboração de um dos indicadores de impacto acadêmico institucional.

d) Aumento significativo de sinergias entre a equipe do projeto e a equipe de TI do Centro. Do ponto de vista de gerenciamento de projetos, a maior melhoria foi o início da implementação de funcionalidades no iN por desenvolvedores da equipe de TI do CGEE. Como principal resultado, destaca-se a elaboração de arquivo descritor JSON para a importação de dados de patentes da base Derwent, da Clarivate Analytics, que permitirá a análise de redes de similaridade semântica e, com algumas adaptações, de redes de coocorrência de códigos de patentes. Também é importante relatar a colaboração entre as equipes na

especificação de requisitos para uma nova rotina de varredura e exclusão de currículos inativos, uma necessidade antiga para melhorias de qualidade da base de dados do Centro, adaptação do iN para a atualização tecnológica para a tecnologia REST dos *web services* do CGEE e a atualização do controle de versionamento do antigo SVN para o modelo git, com migração do repositório das versões armazenadas do programa iN. Essa migração também tornou possível automatizar todo o processo de construção e de liberação do iN, facilitando a manutenção do sistema pelo CGEE. Além disso, a experiência de versionar o iN no git motivou a equipe deste projeto a propor a migração todos os protótipos para este modelo.

e) Desenvolvimento de novas funcionalidades na ferramenta insight Data. Além de tarefas usuais de manutenção corretiva ou adaptativa, um avanço importante a relatar é a criação de formatos de exportação de metadados do repositório do iD, tais como: patentes (USPTO), notícias e documentos da fonte “Base de Teses e Dissertações” (BTD). As exportações são feitas em formatos de dados de descritores existentes do iN, facilitando a realização de análises de redes de similaridade semântica entre esses documentos com o uso de funcionalidades que não existem no iD.

f) Criação de arquivo executável para instalação e configuração automáticas do iNBrowser de acordo com novas restrições do navegador Firefox. Esse empacotamento da ferramenta como executável tornou mais prática a entrega das redes preparadas pelas diversas equipes do CGEE aos clientes finais, pois os conjuntos de redes preparadas são acessados de forma mais simples e direta. Como este foi o único desenvolvimento relacionado ao iNB em 2020, optou-se por não agregar o manual da ferramenta neste relatório. O manual da última versão com as funcionalidades do iNB foi incluído no relatório de 2019.

g) Desenvolvimento de novo método de esparsificação de agrupamentos densos, como os de coautorias e de similaridades semânticas de grandes colaborações científicas. Esse método tem que ser mais testado e é potencialmente inovador para a área de cientometria, pois análises bibliométricas que envolvam grandes colaborações (como em física de

partículas, por exemplo) são tão enviesadas por esses agrupamentos densos que os pesquisadores optam por retirar dados de grandes colaborações da análise.

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, essas ideias foram implementadas em códigos escritos em linguagem Java, para aproveitar a ubiquidade de aplicações existentes, a universalidade de plataformas que podem executar códigos Java e a disponibilidade de ambientes de visualização nessa linguagem que aceleravam o desenvolvimento com a concentração de esforços no desenvolvimento da camada de acesso e tratamento dos dados (*back end*). A ferramenta insightNet, cujo *back end* foi quase todo originalmente desenvolvido no Centro para ser executado aproveitando os recursos visuais da ferramenta Gephi, é o exemplo mais bem-sucedido dessa filosofia de desenvolvimento.

A partir de 2016, novas abordagens foram implementadas para suprir novas demandas que surgiram. Por um lado, tornou-se necessário o tratamento de outras bases de dados textuais. Por outro, as visualizações fornecidas pelo aplicativo Gephi já não eram suficientes, do ponto de vista de acabamento, para as entregas das análises de redes realizadas no Centro por 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 e incômodo 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 tratamento de novas bases de dados textuais, particularmente notícias, o CGEE contratou e colaborou no desenvolvimento da plataforma insight Data, que é baseada em várias linguagens de programação e contém uma camada de visualização (*front end*) desenvolvida em JavaScript especificamente para essa aplicação.

Para o modelo de entrega de produtos com *front end* de análise visual interativa de dados foi desenvolvida internamente a ferramenta insightNet Browser. 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 consideravelmente melhorada 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 (provida pelo CNPq) 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 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. Esta última etapa do processo deverá ser iniciada em 2021.

Em 2020, várias aplicações foram desenvolvidas de acordo com essa nova filosofia. No que segue, serão resumidos apenas 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. Os casos nos quais já houve resposta de usuários serão destacados.

a) Nova versão do sumarizador semiautomático de textos. A versão inicial desse aplicativo foi idealizada para otimizar a análise de textos longos, degravações de reuniões e entrevistas e foi empregada, ainda em 2019, para contribuir nas análises da Política Nacional de Inovação. Em 2020, algumas otimizações foram realizadas e os casos de uso foram estendidos para contribuir na elaboração de resumos executivos de relatórios. A partir de uma ideia inovadora sugerida por

usuários do projeto “Observatório de CTI”, serão realizados em 2021 testes de redes de similaridade semântica entre textos com tamanhos normalizados pelo sumarizador. Essa ideia pode ajudar a resolver o problema de disparidade de tamanhos de textos, comum em processamento de linguagem natural (o conteúdo semântico de textos maiores tende a se sobrepor aos conteúdos de textos menores). É importante ressaltar que essa ferramenta é capaz de resumir textos na maioria das línguas e a inserção de dados pode ser feita seja por leitura do arquivo de texto, seja pela ferramenta de cortar e colar dos sistemas operacionais. Um exemplo de tela do sumarizador segue abaixo.

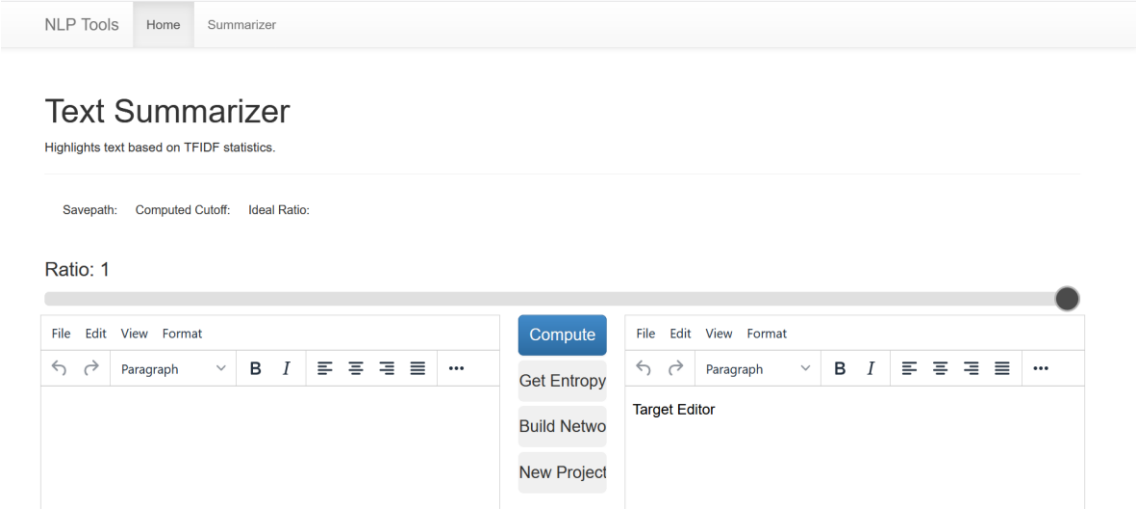


Figura 1: Tela de entrada do sumarizador de textos.

b) Aplicativo para explicitação de termos conectores em redes de similaridade semântica. 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. Para caracterizar o conteúdo temático do *cluster*, as ferramentas de análise desenvolvidas coletam as palavras-chave dos textos e um escore de frequência dessas palavras-chave quase sempre é suficiente para a sua classificação temática. Entretanto, em conjuntos de dados

que não contém 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 coleta de termos conectores visa extrair o grau de relevância das palavras, quantificando as suas participações nas composições dos pesos das arestas. 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. O protótipo dessa ferramenta, cujos conceitos foram integralmente idealizados no CGEE, foi testado na análise da consulta pública da Estratégia Nacional de Inovação e uma variação do seu algoritmo foi implementada em análises de grandes volumes de resumos de artigos realizada pelo Observatório de CTI. Um exemplo da visualização fornecida pela ferramenta é exibido na Fig. 2.

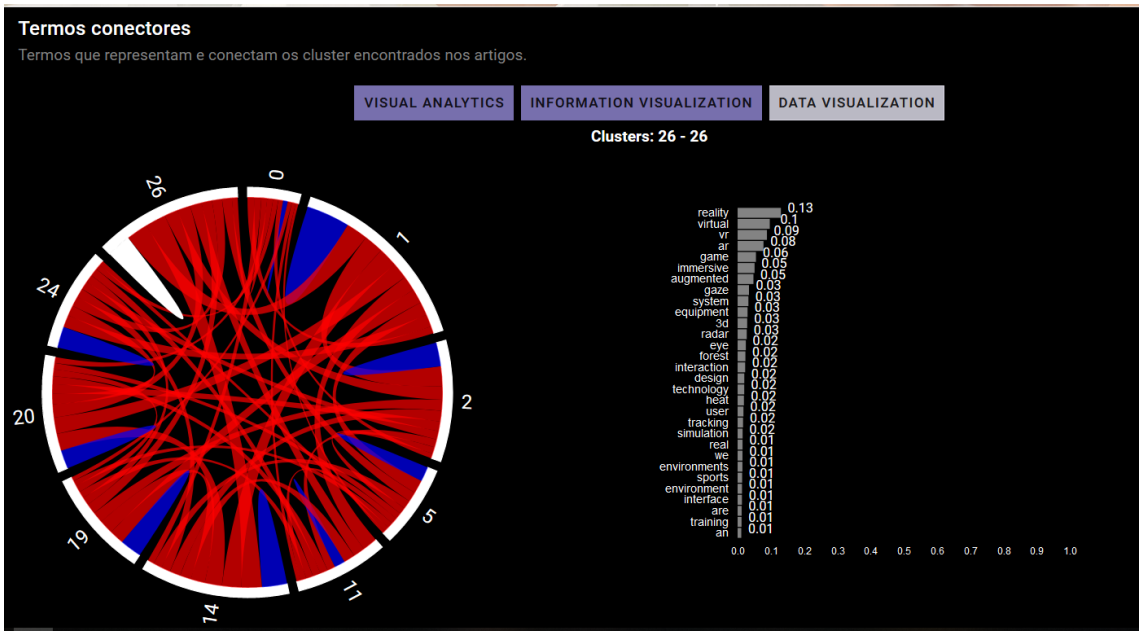


Figura 2: Visualização de termos conectores do *cluster* 26 (ressaltado em branco na imagem da esquerda) de um conjunto de resumos de artigos baixados no Web of Science com o termo de busca “data visualization”. Os escores dos termos conectores exibidos à direita permitem inferir que o *cluster* trata de visulização de dados em contextos de realidade virtual.

c) Visualizador de mapas de calor de coocorrências, correlações e probabilidades condicionais entre dados de planilhas. Concebida inicialmente para agregar informação em análises de dados de consultas públicas realizadas na plataforma insightSurvey, esse protótipo permite quantificar em uma matriz os graus de correlação e de probabilidades condicionais evidenciados por intensidade de cor entre metadados numéricos ou de respostas de itens (mapas de calor). Por exemplo, o usuário pode avaliar, dada a quantidade de respostas em um item da pesquisa, quais são as probabilidades de ocorrência de cada uma das demais respostas apenas inspecionando a linha ou coluna correspondente. O protótipo foi empregado na análise da consulta pública sobre a Estratégia Nacional de Inovação. Posteriormente, as funcionalidades da ferramenta foram estendidas para quaisquer análises de metadados tabulados em planilhas e essa nova versão está em fase de validação interna por parte da equipe do projeto. As Fig. 3a e 3b, abaixo, mostram duas das quatro etapas da análise de dados, a etapa de seleção e de escolha do tipo de variáveis a serem consideradas e a visualização de coocorrências triplas em um gráfico 3D, respectivamente.

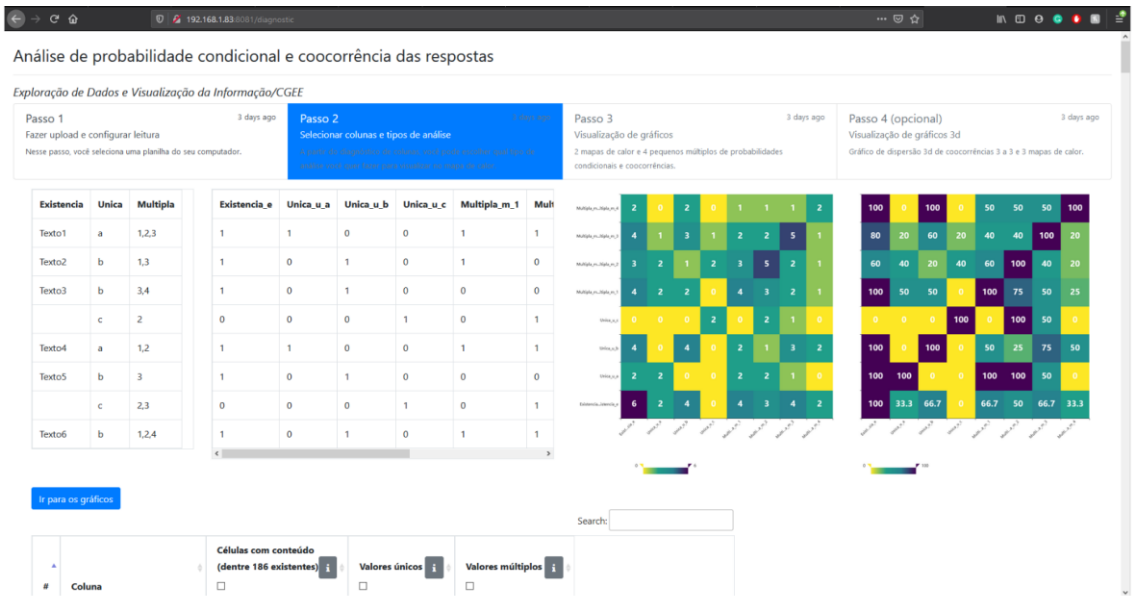


Figura 3a: Tela da ferramenta para geração de mapas de calor com as opções de escolha de variáveis e de tipos de contagens a serem consideradas na análise, após o arquivo de dados (por

exemplo, uma planilha Excel) ter sido carregado.

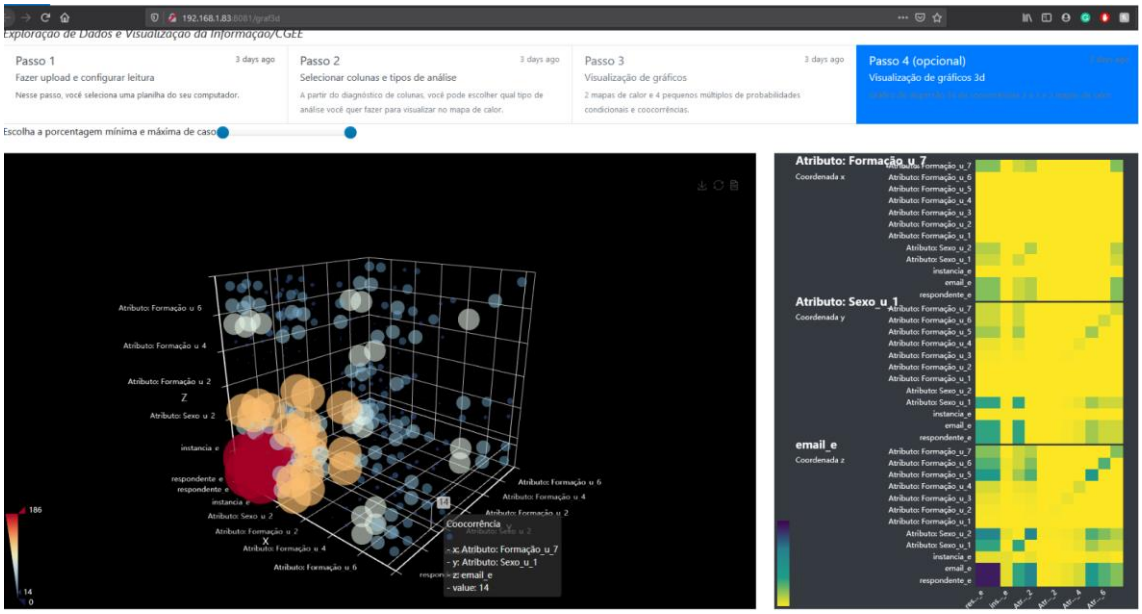


Figura 3b: Tela com os resultados dos cruzamentos de dados três a três (neste caso, de coocorrências). A imagem da direita é o mapa de calor com vermelho se referindo a o número maior de coocorrências por tripleto e, a esqueda, podem ser vistas as projeções da matriz 3D.

d) Internalização de protótipo com adaptações e melhorias do *Science Overlay Map*. O *Science Overlay Map* (SOM) é um programa no qual é possível distribuir visualmente um conjunto de artigos em uma rede de coocorrências entre áreas de pesquisa de acordo com suas classificações no Web of Science. Trata-se de um recurso bastante empregado na área de bibliometria que foi idealizado pelo pesquisador Loet Leydesdorff, da Universidade de Amsterdã. SOMs são particularmente úteis para caracterizar evoluções de publicações de pesquisadores multidisciplinares, instituições ou países entre áreas de pesquisa. A versão básica da aplicação é disponibilizada na página pessoal do pesquisador, mas a equipe do projeto realizou diversas melhorias na ferramenta original, incluindo vários recursos para ressaltar subredes, inclusão de termos de busca e exibição de estatísticas de frequência de artigos por área. Essa versão foi disponibilizada em uma página de acesso interno do Centro, para testes e validação. A intenção é de, caso a aplicação seja validada, desenvolver uma

versão mais robusta e configurável que carregue os dados bibliométricos de um dado conjunto de artigos e gere automaticamente seu respectivo SOM. Um exemplo de SOM, incluído no ambiente de análise visual de dados construído para exibir um levantamento do estado da arte da visualização de dados, segue abaixo (Fig. 4).

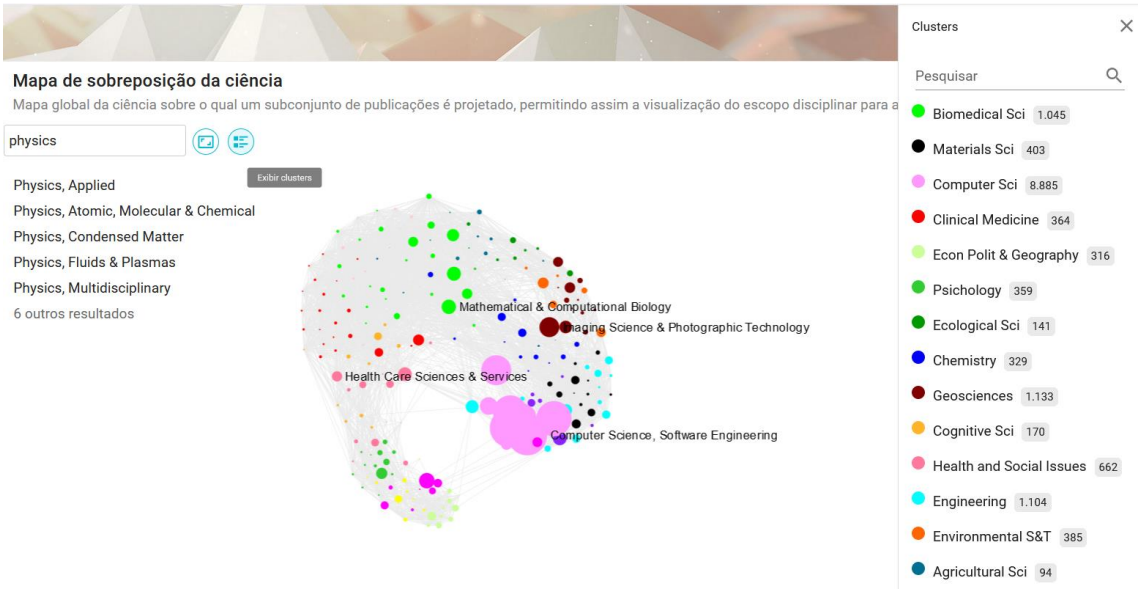


Figura 4: Versão de Science Overlay Map com diversos recursos de navegação e busca desenvolvida no CGEE. A versão original é de autoria de Loet Leydesdorff e está disponível na página https://www.leydesdorff.net/.

e) Internalização de protótipo com adaptações e melhorias do "Seealsology". O Seealsology é um software livre criado pelo Médialab, da *Fondation Nationale des Sciences Politiques*, para mapeamento de verbetes da Wikipedia. Embora a Wikipedia não seja a melhor fonte de informações sobre um determinado tema, frequentemente seus artigos podem ser lidos para o usuário ter uma visão geral rápida a seu respeito. O Seealsology usa os links da seção "*See also" para* gerar automaticamente redes desse tipo de citação, onde as arestas se referem às citações e os nós, aos artigos. As redes produzidas dão uma excelente visão geral do tema e de suas relações com temas correlatos. A ferramenta também possibilita ao usuário acessar o artigo original na página da Wikipedia. A versão

desenvolvida no CGEE retém todas as funcionalidades da versão do Médialab e acrescenta resumos dos artigos para uma leitura rápida e uma ferramenta que torna as buscas muito mais convenientes. Um exemplo de tela da ferramenta, incluída no ambiente de análise visual de dados construído para exibir um levantamento do estado da arte da visualização de dados, é mostrado na Figura abaixo.

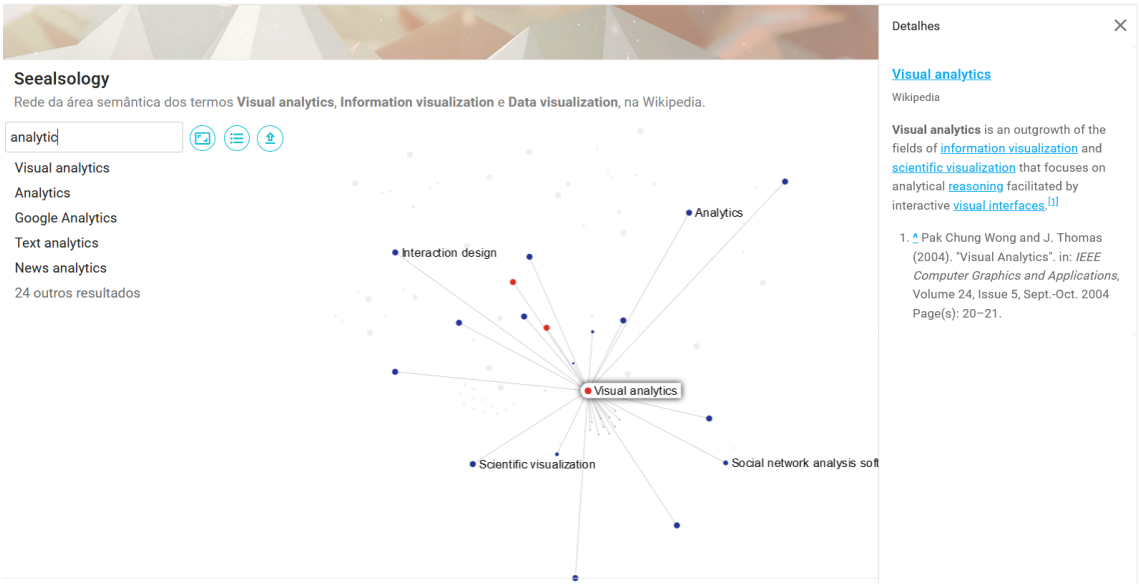


Figura 5: Versão da ferramenta “Seealsology” com novas funcionalidades desenvolvidas no CGEE. A versão original do programa pode ser explorada na página https://densitydesign.github.io/strumentalia-Seealsology/

f) Editor de arquivos XML de currículos Lattes. Os currículos da plataforma Lattes são disponibilizados no formato XML, que são arquivos de texto cuja estrutura é delineada por códigos (*tags*) que identificam e localizam conteúdos informados pelo usuário da plataforma. A ferramenta permite editar conteúdos de *tags* existentes de um bloco inteiro de currículos, alterando-os com a inclusão de dados recolhidos de uma planilha. No caso de uso real trabalhado, num trabalho de caracterização bibliométrica das unidades de pesquisa (UPs) do MCTI, as UPs enviaram para o CGEE os dados de períodos em que pesquisadores servidores, visitantes e estudantes tiveram vínculo com as

instituições. Esses dados foram inseridos nos currículos na *tag* “Ensino fundamental”, que normalmente não é empregada. Após adaptação no iN, os novos conteúdos das *tags* editadas foram lidos e usados para filtrar apenas as publicações dos períodos de vigência dos vínculos das pessoas com as instituições. Estendida para editar outros tipos de arquivos XML além do Lattes, a ferramenta tem um potencial promissor de uso como coadjuvante em análises de impacto de políticas públicas efetivadas em intervalos específicos de tempo, por exemplo. O maior ganho esperado é que o processo de extração, tratamento e normalização de dados se torne automatizado e facilitado sem a necessidade de o usuário entender da programação subjacente, de modo que análises exploratórias possam ser feitas por usuários leigos na programação, mas experientes nos assuntos estudados. A figura a seguir mostra a tela de entrada de dados da ferramenta.

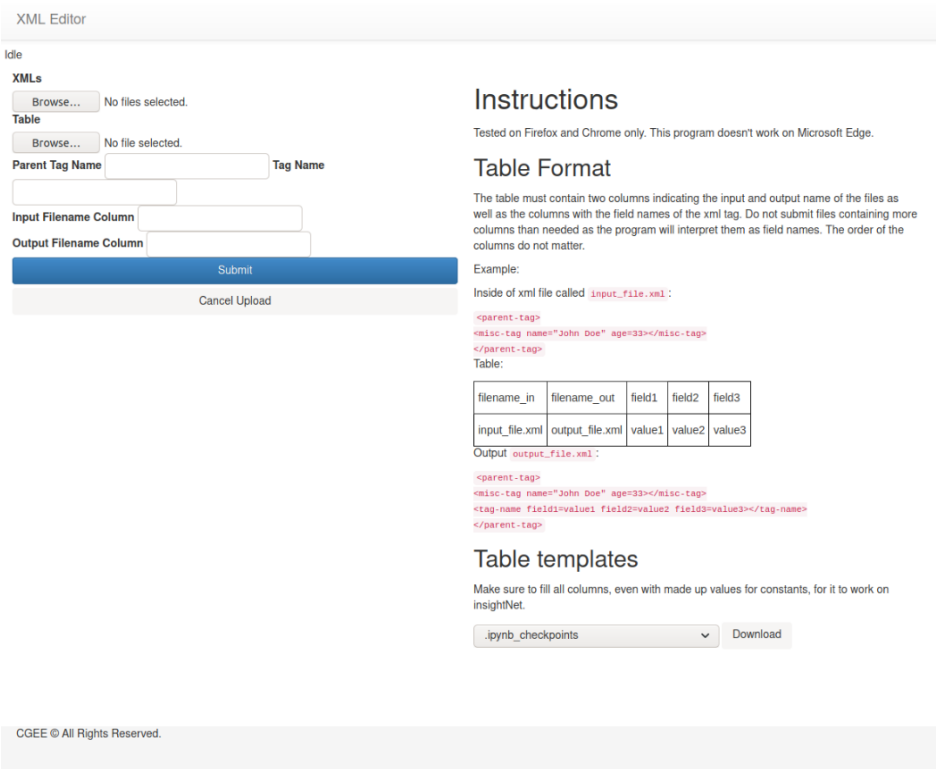


Figura 6: Tela de abertura do editor de *tags* de XMLs. Note-se que há duas entradas de dados: a da pasta contendo os arquivos XML a serem editados e o arquivo de metadados a serem importados para estes arquivos. As demais opções servem para escolher os nomes das *tags* a

serem editadas e nomes de arquivos de saída.

g) Implementação e testes de várias alternativas de visualização baseadas nas bibliotecas D3 e "echarts", em Javascript, algumas integradas no ambiente de análise visual de dados descrito anteriormente. Como exercício para futuros projetos que explorem a dimensão temporal, bem como conceitos de inte-ratividade na visualização, foi testada uma animação do globo terrestre com representações da produção acadêmica dos vários países, ainda com a base de metadados de artigos sobre visualização de dados e suas coautorias, como visto na Fig. 7 a seguir. Na versão web dessa figura, o globo gira em torno de um eixo fixo, mas o usuário pose girá-lo em torno do eixo que desejar usando o cursor, além de dar zoom e realizar buscas nos metadados.

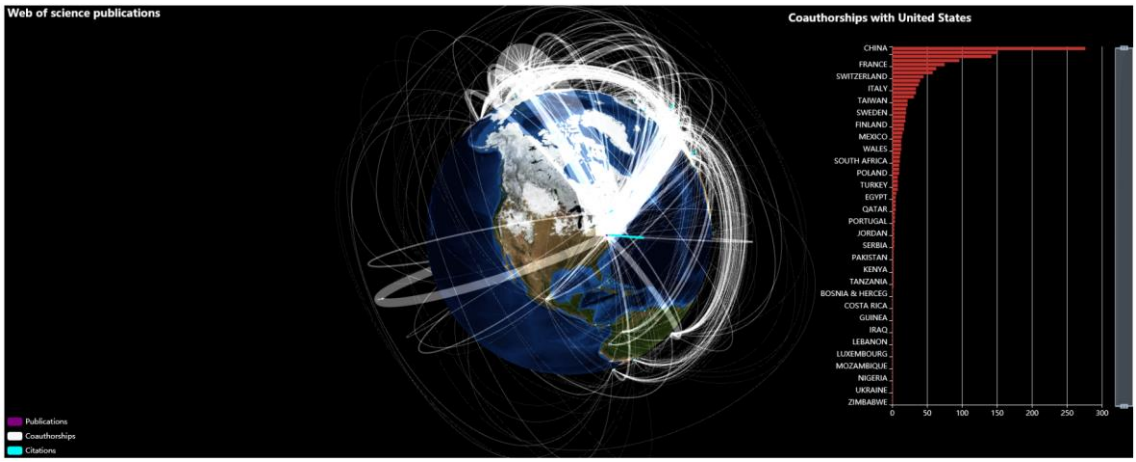


Figura 7: Imagem da animação do globo terrestre evidenciando com barras de cores diferentes o volume de publicações e de citações de um determinado país. As espessuras das arestas curvas no globo e o gráfico de barras representam o volume de coautorias de um país selecionado, no caso em destaque, os EUA.

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 ferramen-tas com foco na prototipagem rápida das ideias e algoritmos discutidos pela equipe do projeto. Dentre as iniciativas realizadas, destacam-se:

a) Levantamento de bases de dados e algoritmos existentes no CGEE da perspectiva dos usuários, incluindo suas descrições e exemplos de uso. Essa iniciativa, reportada em produto específico, 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, provavelmente a melhor alternativa para a exibição do organograma seja transpô-lo para o formato HTML e hospedá-lo em uma página na intranet do Centro.

b) Elaboração de 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 o fluxo descrito na introdução da Seção 2 deste relatório, 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, e integração dos códigos validados com interfaces web que facilitam o uso e testes por parte de usuários das assessorias;

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 nova 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 com sequência 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 buscar 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 visavam usualmente competir com soluções existentes no mercado, mas também não excluia a possibilidade de testá-las para descartá-las, simplesmente usá-las (como exemplo, o uso do Tableau como ferramenta de prototipação de visualizações de dados) ou adaptá-las a objetivos específicos do Centro (como exemplos, pode-se citar os SOMs de Leydesdorff e

o Seealsology descritos acima). Como consequência lógica dessa diretriz, o grupo tendeu naturalmente a empregar quase que exclusivamente soluções de código aberto.

Uma 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, por enquanto, 6 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) Uso de algoritmos de aprendizado de máquina para a geração de modelos de classificação de textos curtos. Modelos desse tipo foram treinados para classificar perguntas e queixas de usuários do insightSurvey e foram testados em consultas públicas do MEC e do MCTI.

b) Emprego de equações diferenciais da epidemiologia (SEIR) para a caracterização de dinâmica de redes complexas. Neste caso, a meta foi examinar a viabilidade de modelos SEIR para a caracterização de sinais emergentes em conjuntos de textos a partir da dinâmica de surgimento de arestas.

c) Justificação teórica de candidatos a "cutoff" de pesos irrelevantes (esparsificação do grafo, redução dimensional ou, ainda, network backbone extraction, nos jargões técnicos de diferentes subáreas) para otimização informacional/estatística de redes de similaridade semântica. Essa discussão não foi conclusiva e teve que ser suspensa para o grupo abordar outras prioridades, mas terá que ser retomada em 2021.

d) Análises de algoritmos eficientes para processamento de linguagem natural em Python para grandes volumes de documentos (tipicamente 20-30 vezes mais documentos do que a atual capacidade instalada nas ferramentas de uso comum no centro, como o insghtNet). Neste caso foram discutidas soluções empregadas em um importante trabalho de caracterização de centenas de milhares de artigos de pesquisadores brasileiros, realizado no âmbito do projeto

OCTI.

e) Discussão sobre modelos de linguagem, com ênfase do novíssimo GPT3, lançado em meados de 2020. Essa discussão foi iniciada em novembro, motivada por uma apresentação no Journal Club do Centro e deve ser continuada ao longo de 2021, uma vez que, devido aos padrões linguísticos que têm que ser reconhecidos para seu funcionamento, modelos de linguagem têm conotações muito mais abrangentes do que apenas a geração de texto que é divulgada.

f) Estudos sobre a API WebGL, do JavaScript, que utiliza recursos da unidade de processamento gráfico da placa de vídeo do computador para renderizar gráficos interativos em 2D e 3D. Essa API foi posteriormente empregada nas implementações do globo 3D, na exibição de redes 3D do projeto OCTI e na versão de exibição em grafos 3D da ferramenta de coleta de termos conectores. Avanços no uso desse tipo de tecnologia possibilitará, em breve, a exibição de grandes quantidades de dados, em ambiente web, com boa performance (por exemplo: redes com mais de 100 mil nós e outras visualizações 3D).

Apêndice A: Manual CGEE Insight Net 3.2.6

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 exploradas. Este guia de usuário descreve a versão 3 do *plug-in*, com detalhes sobre a instalação dos seus componentes e suas principais funcionalidades.

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*:

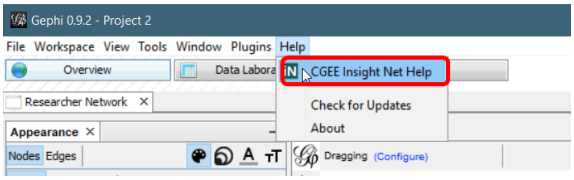


Figura 1.1 Opção para abrir a ajuda on-line

1.3 Funcionalidades experimentais

O CGEE Insight Net está em um processo de aprimoramento constante e a grande maioria das funcionalidades 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 específicos. 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:

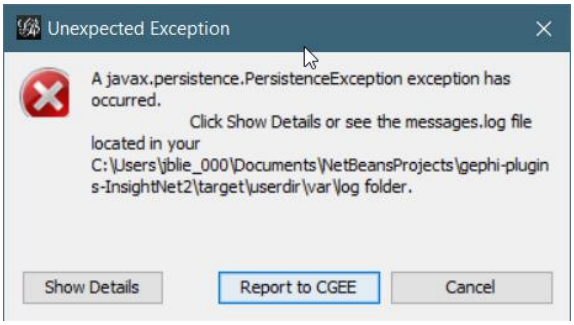


Figura 1.2 Mensagem de erros inesperados

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

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

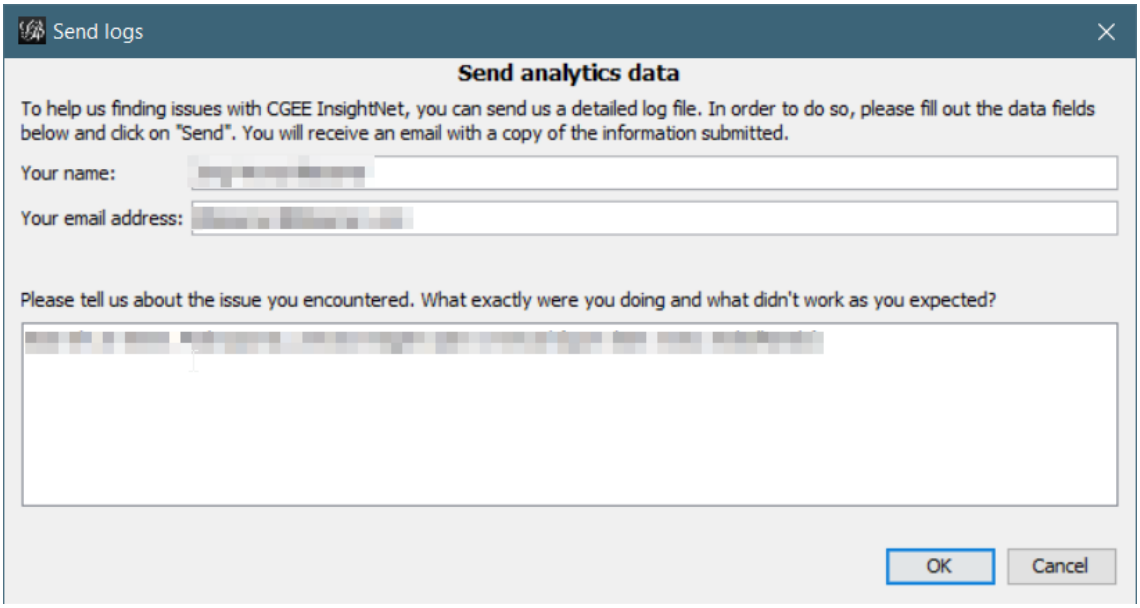


Figura 1.3 Diálogo de envio do protocolo de execução

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:

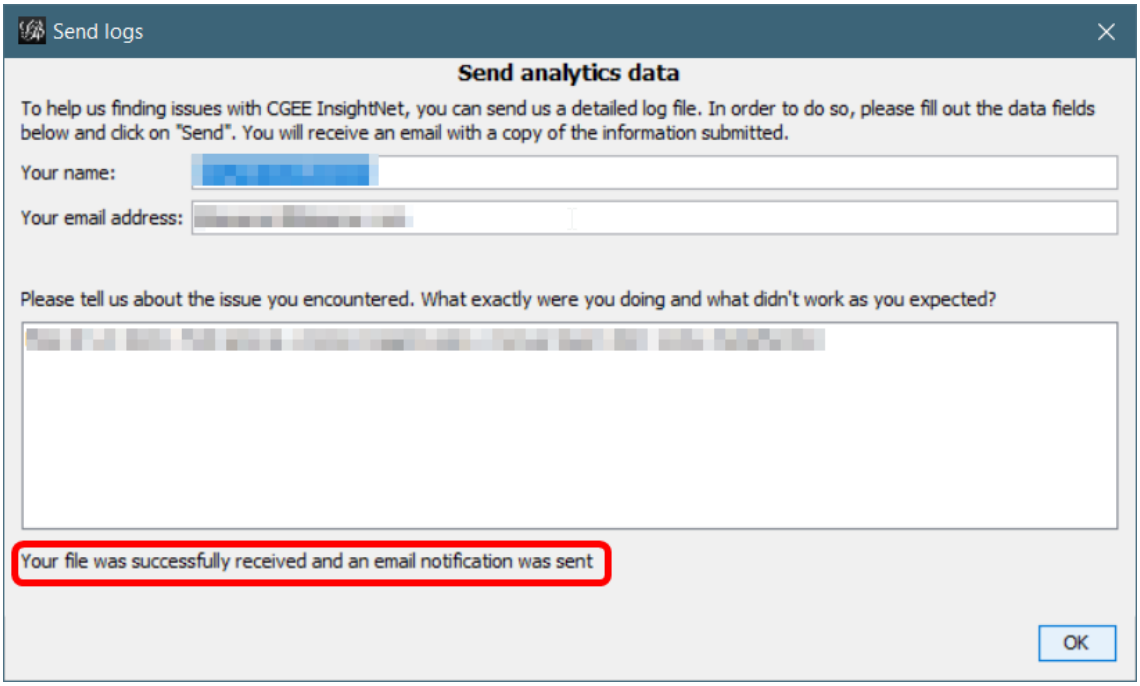


Figura 1.4 Conclusão do envio do protocolo de execução

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:

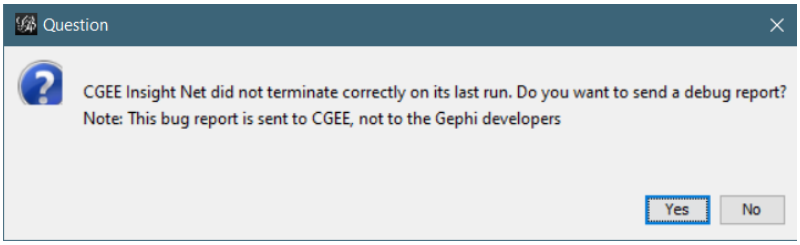


Figura 1.5 Mensagem após o encerramento irregular do Gephi

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

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 [2]

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 configuração 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**.

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*":

[2] https://gephi.org

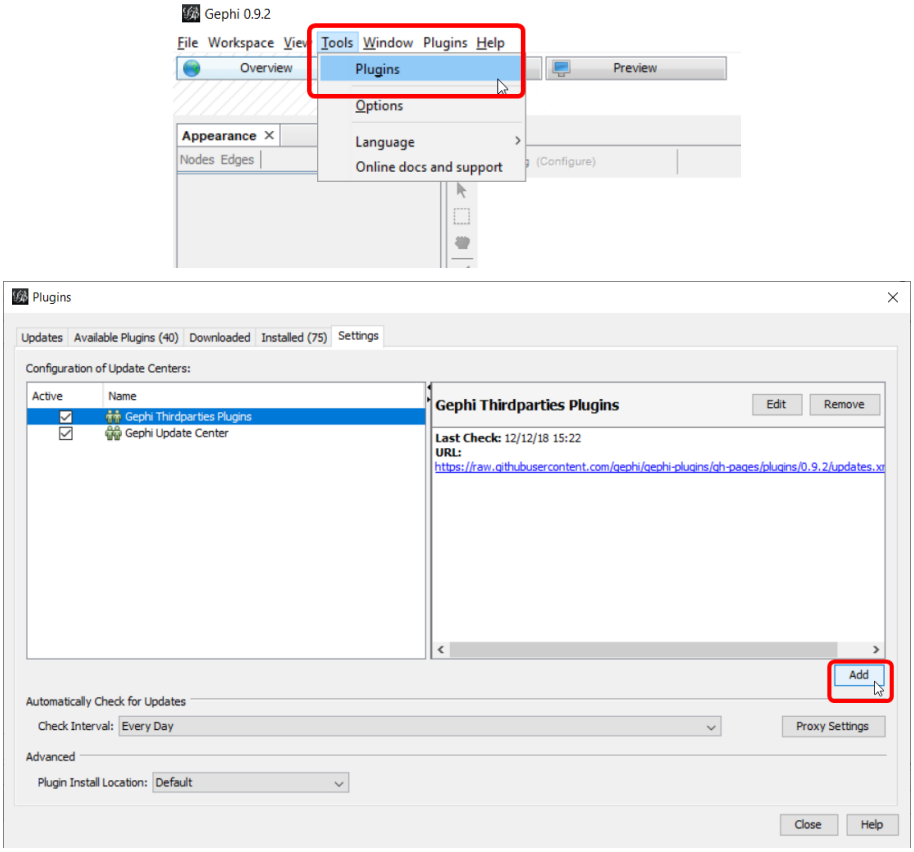


Figura 2.1 Configuração dos centrais de atualização Gephi

No diálogo que aparece, os seguintes dados devem ser especificados:

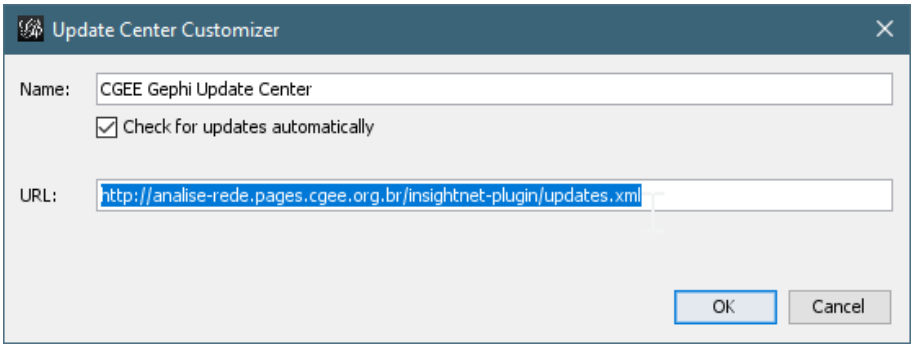


Figura 2.2 Configuração da central de atualizações do CGEE Insight Net

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*":

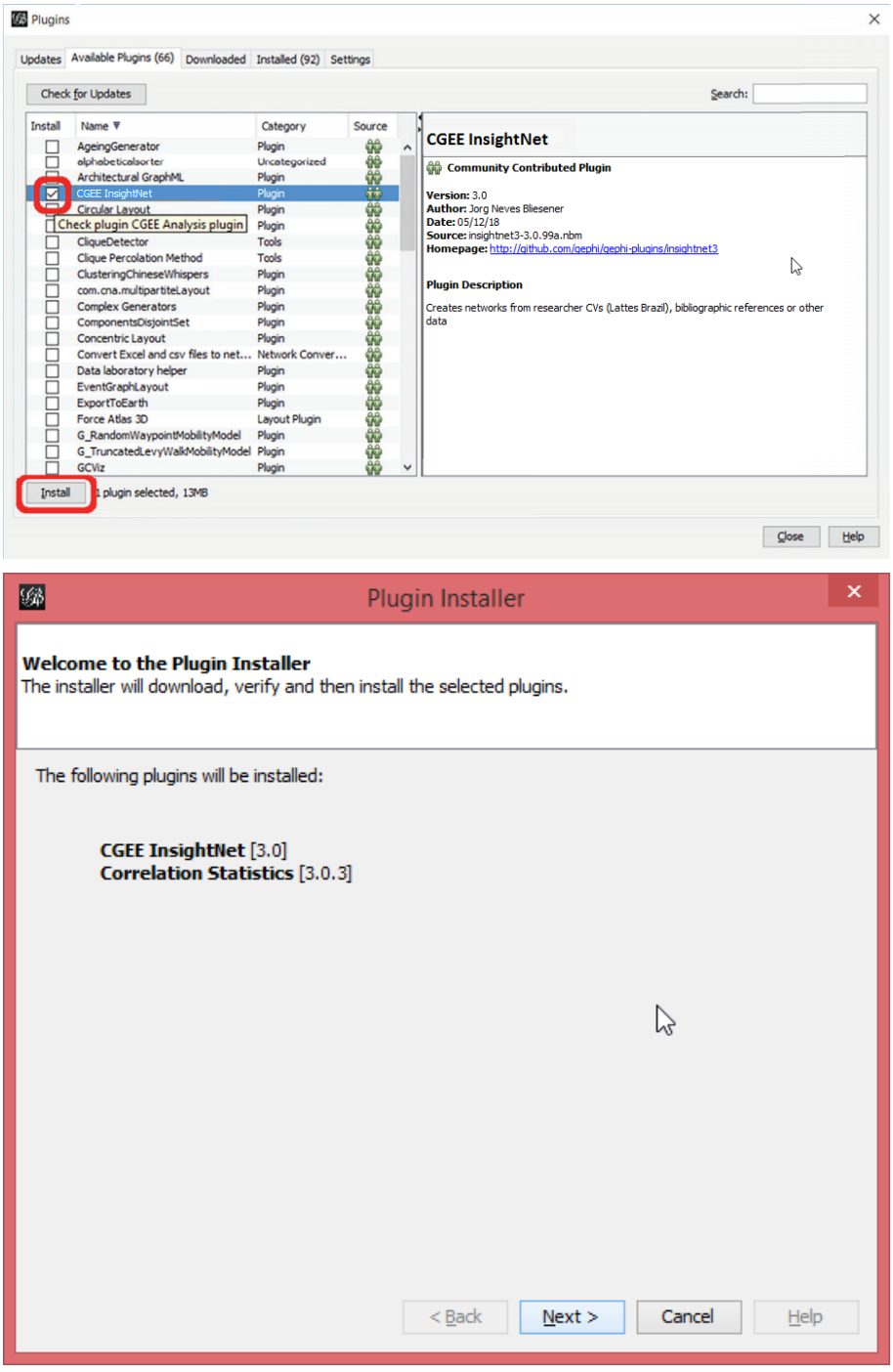


Figura 2.3 Instalação do CGEE Insight Net

Ao chegar à tela de licença, o usuário deve aceitar o texto da(s) licenças exibida(s) para concluir a instalação [3] :

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

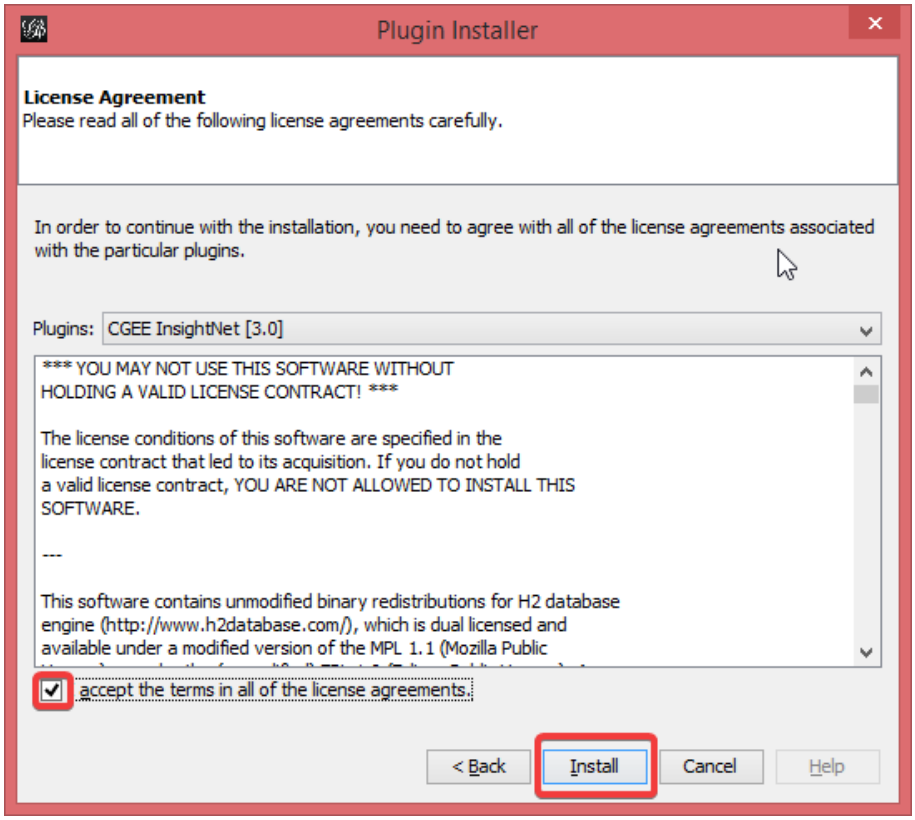
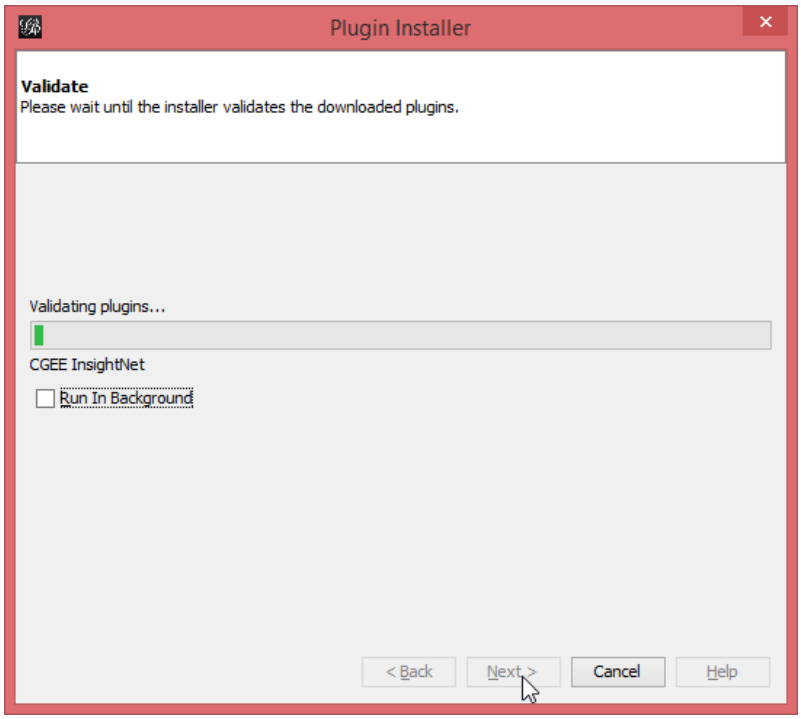


Figura 2.4 Concordância com a(s) licença(s) do produto

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



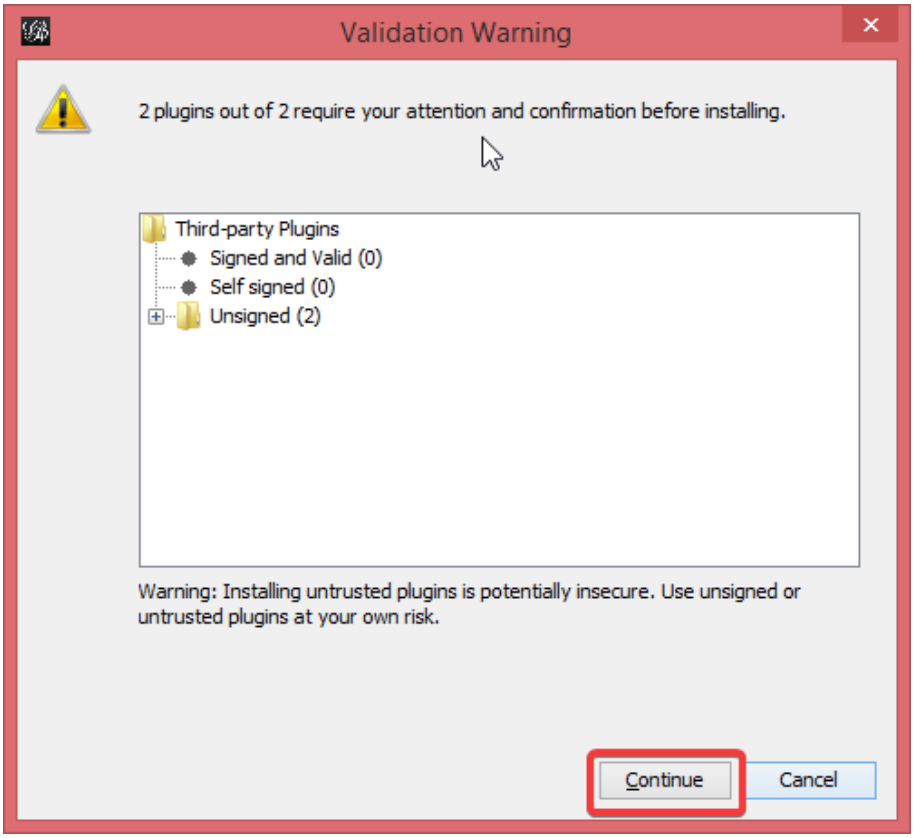


Figura 2.5 Download e aviso de falta de assinatura

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

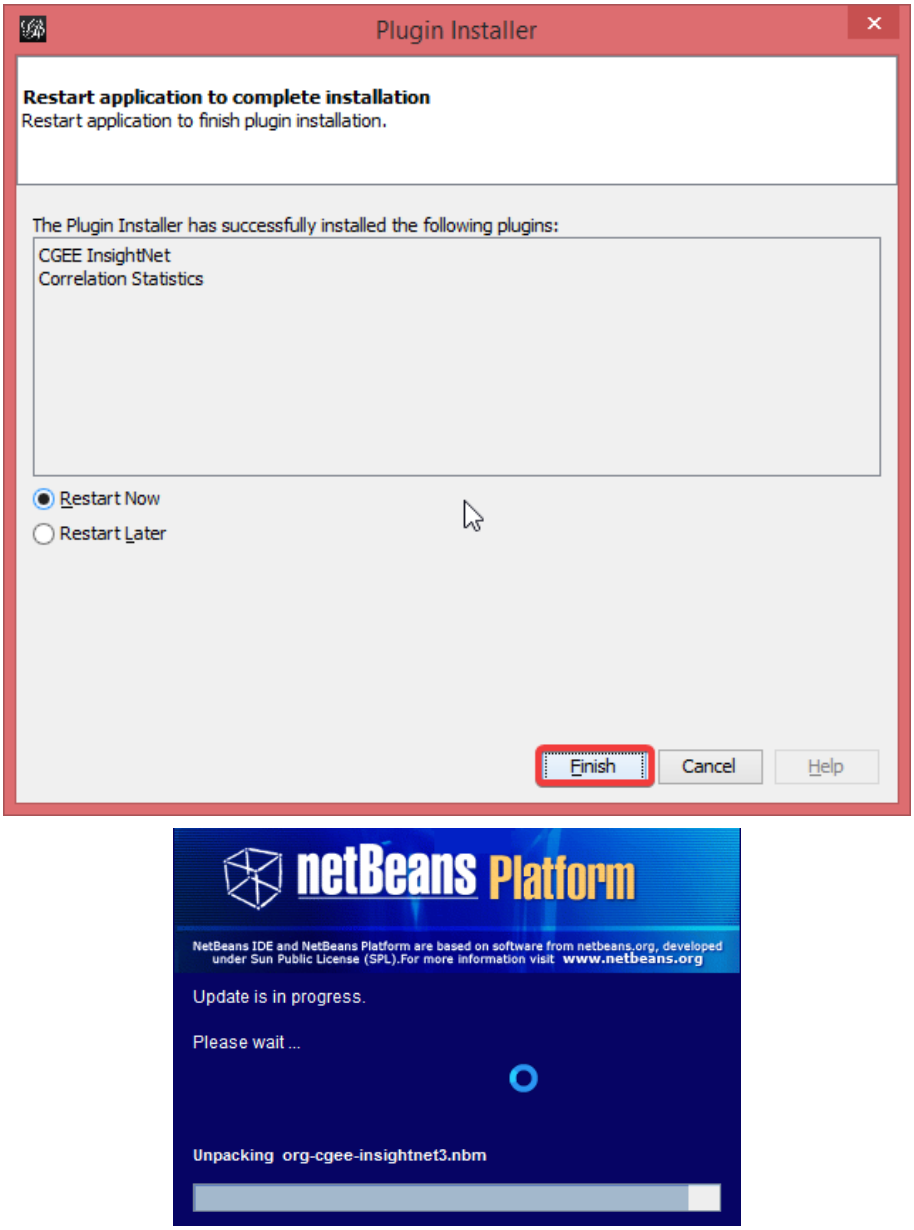


Figura 2.6 Instalação do CGEE Insight Net

Depois de reiniciar, o *GEE Insight Net* tenta alterar o arquivo `gephi.conf` que consta no subdiretório `etc` da instalação do Gephi [4]. Caso essa tentativa tenha êxito, a seguinte mensagem é exibida e o sistema é reiniciado outra vez:

[4] 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:

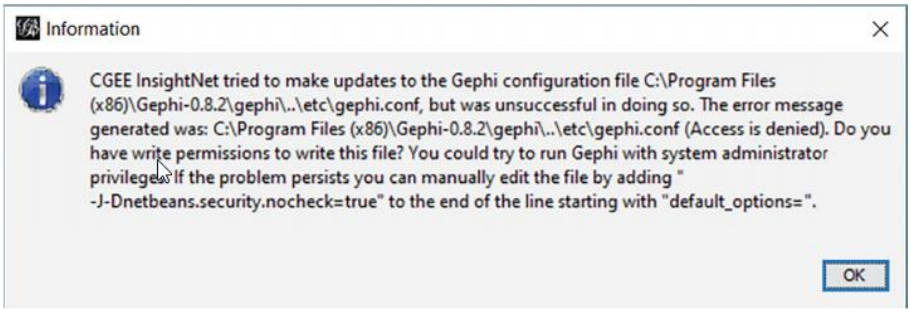
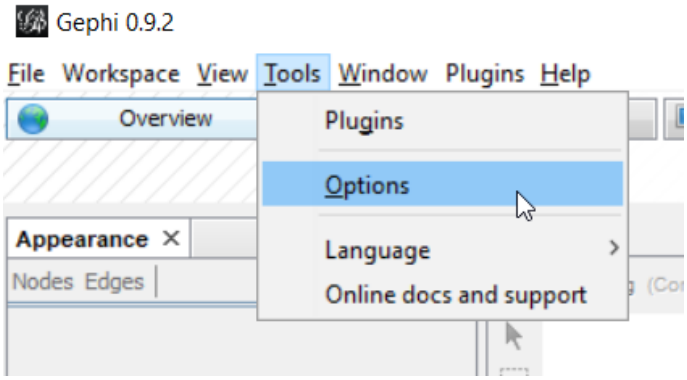


Figura 2.8 Mensagem de alteração malsucedida do arquivo gephi.conf

Nesse caso, a gestão e o uso de licenças adicionais (ver Seção 3.7) fica indisponível e uma das alterações 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 efetivamente, as licenças correspondentes aos módulos do contratados ainda devem ser instaladas conforme descrito na Seção 3.7.



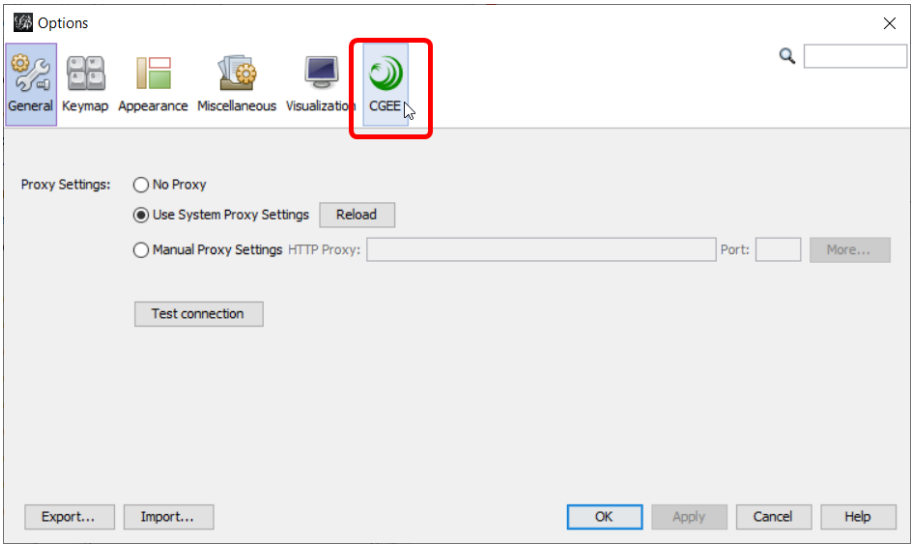


Figura 2.9 CGEE Insight net instalado

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*:

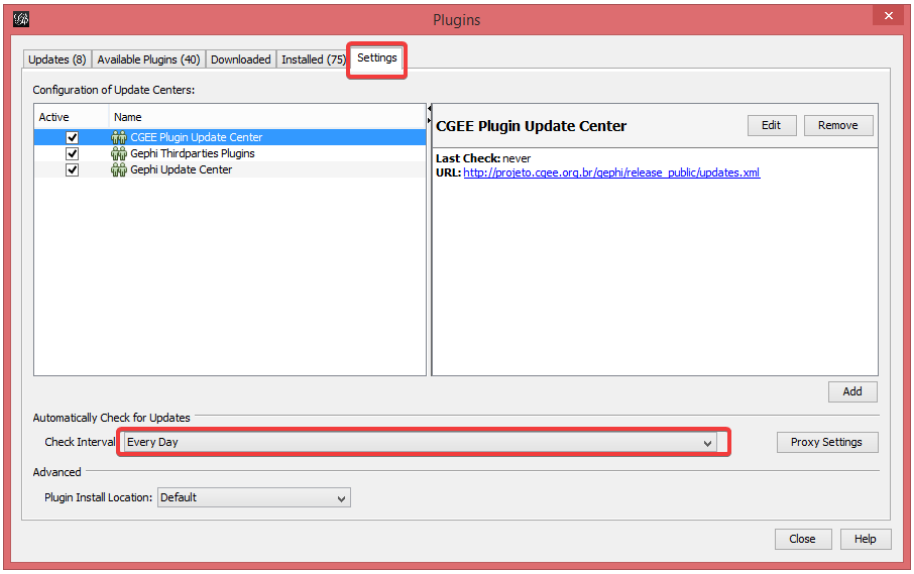


Figura 2.10 Configuração da atualização automática

Caso haja atualizações do *CGEE Insight Net*, um aviso aparecerá no Gephi:

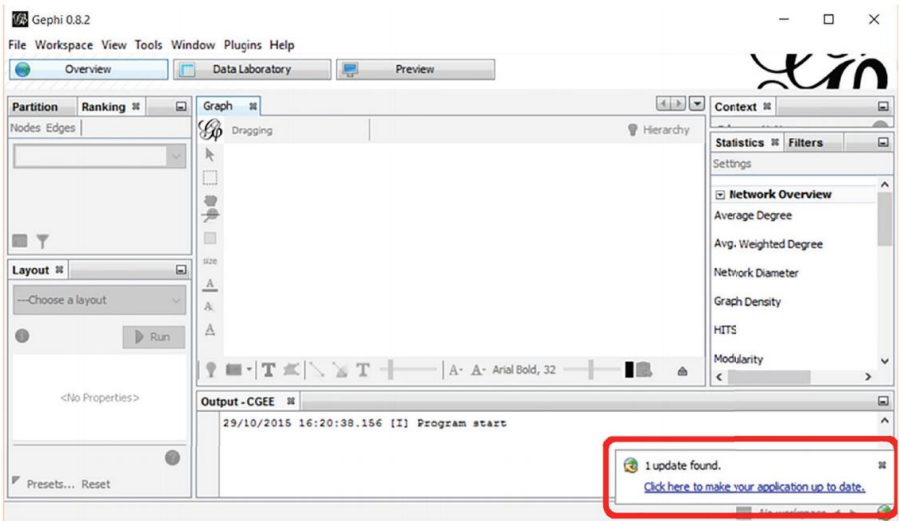
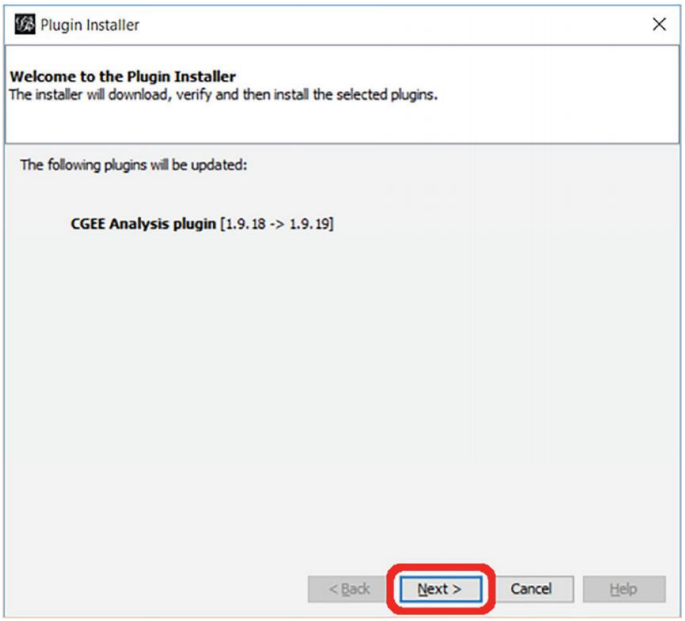


Figura 2.11 Aviso de atualização

Clicando na mensagem de atualização, o Gephi exibe a lista de atualizações disponíveis e, clicando em *Next*, inicia o processo de atualização:



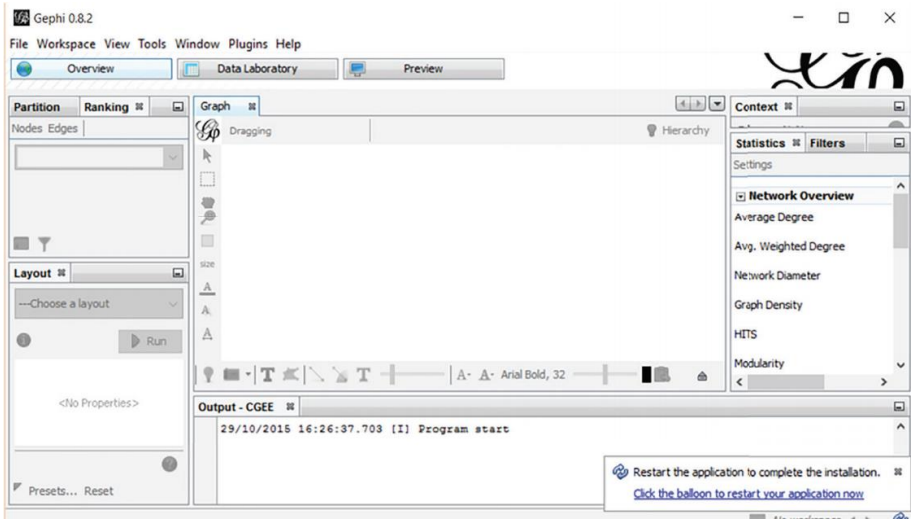


Figura 2.12 Aviso de atualização

Clicando no aviso, o Gephi conclui a atualização e reinicia o programa.

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*.

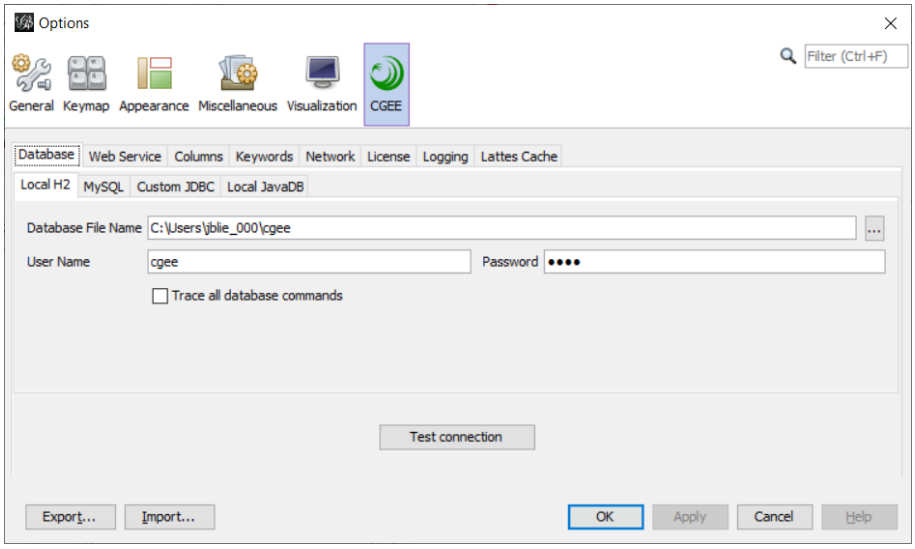


Figura 3.1 Opções de configuração

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:

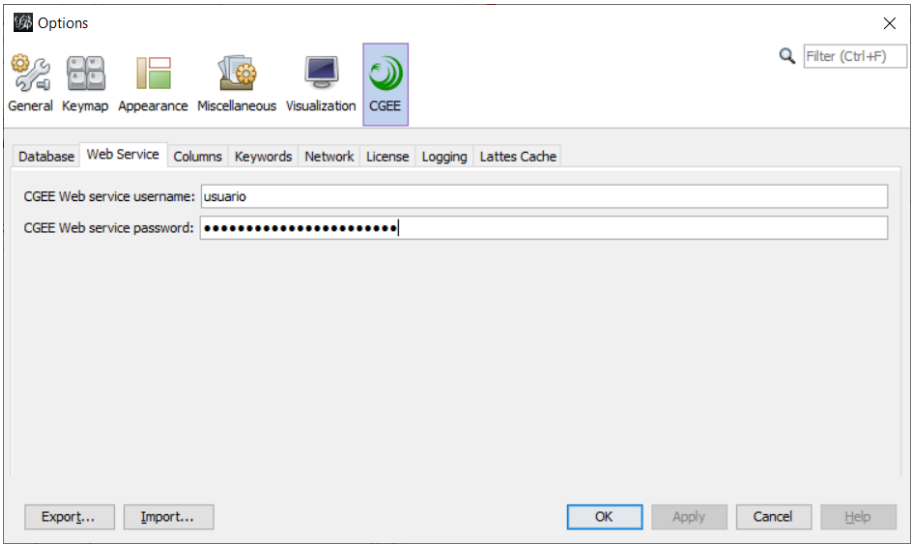


Figura 3.2 Configuração do usuário e da senha para acessar o banco de dados de Currículos Lattes do CGEE

3.3 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 correspondentes 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:

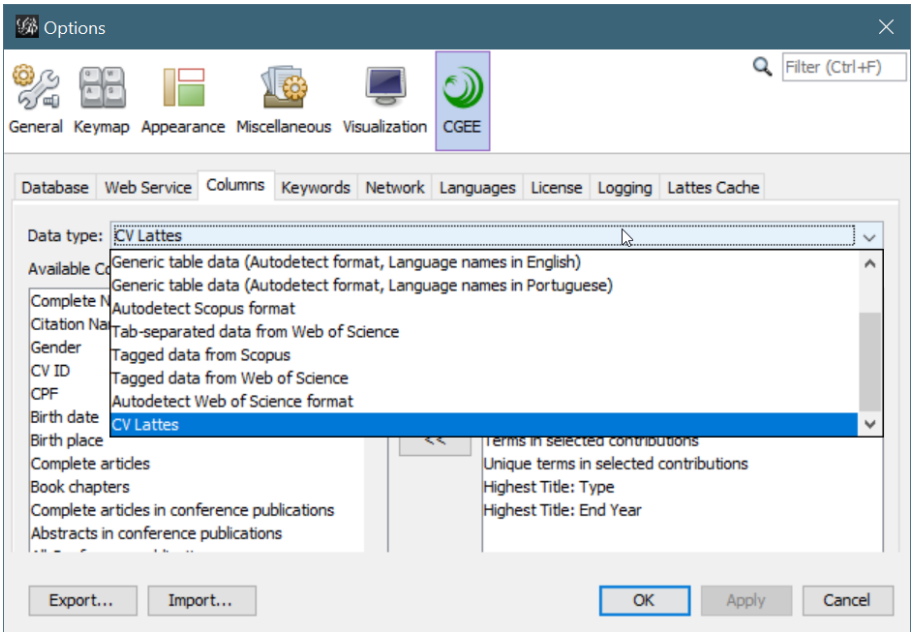


Figura 3.3 Seleção do tipo de dados para configurar as colunas que serão exibidas por padrão

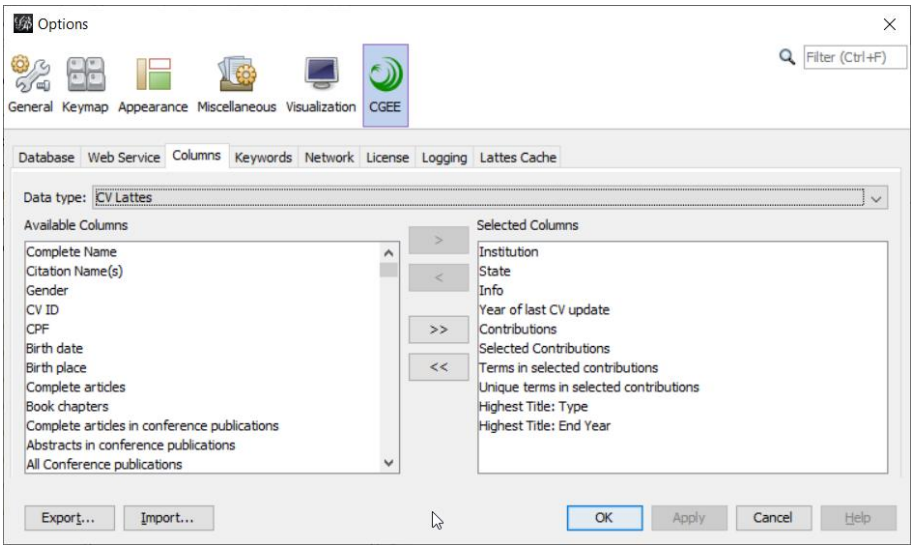


Figura 3.4 Configuração das colunas dos currículos Lattes que serão exibidas por padrão

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

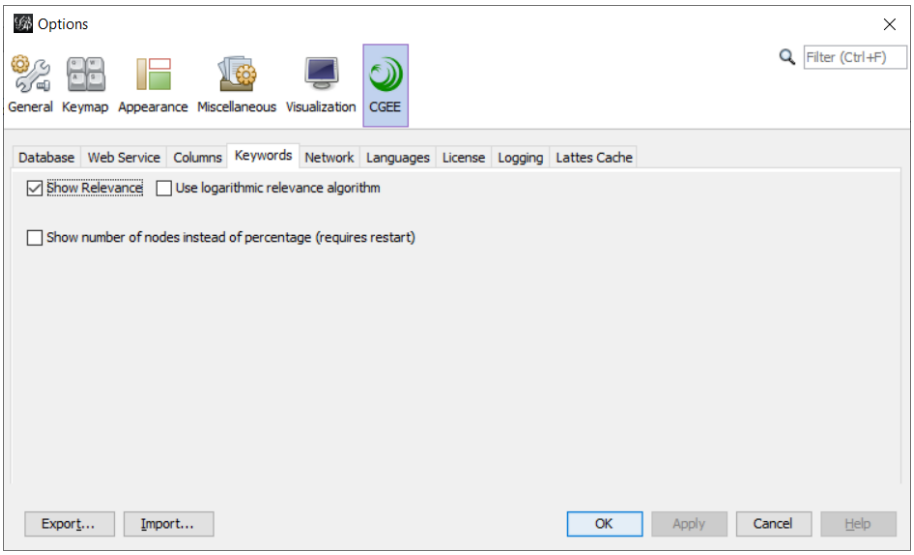


Figura 3.5 Configuração da exibição das palavras-chave

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 logaritmicos 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

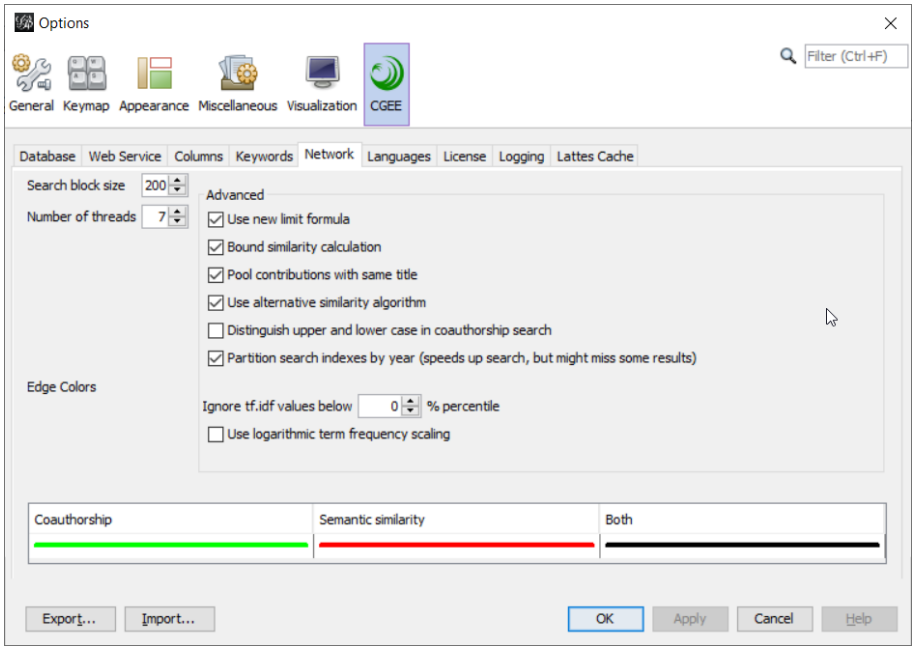


Figura 3.6 Configuração do cálculo e da exibição das redes

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 computador. 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 processamento 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 similaridade. 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 similaridade.
- Geralmente, a pesquisa não distingue entre letras minúsculas e letras maiúsculas e

considera palavras 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 permite 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:

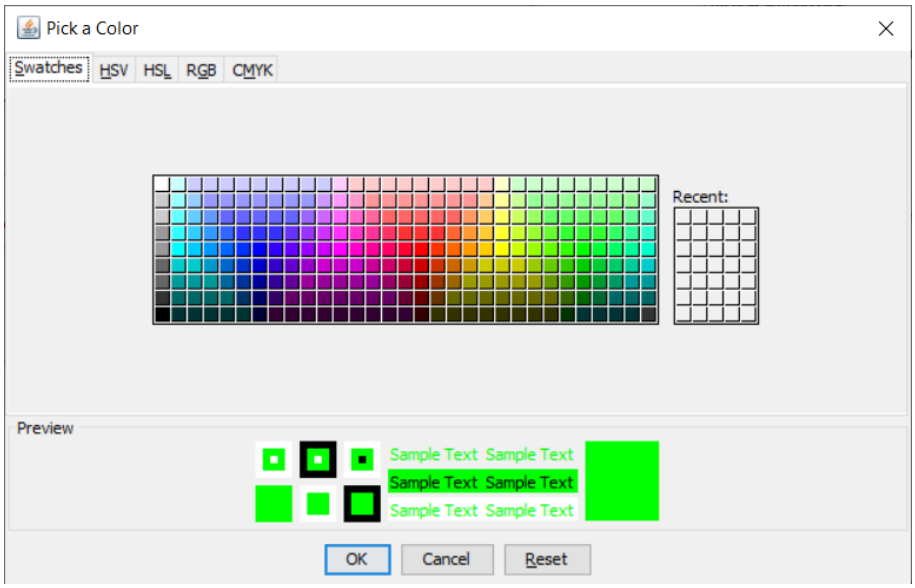


Figura 3.7 Seleção de cores

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*".

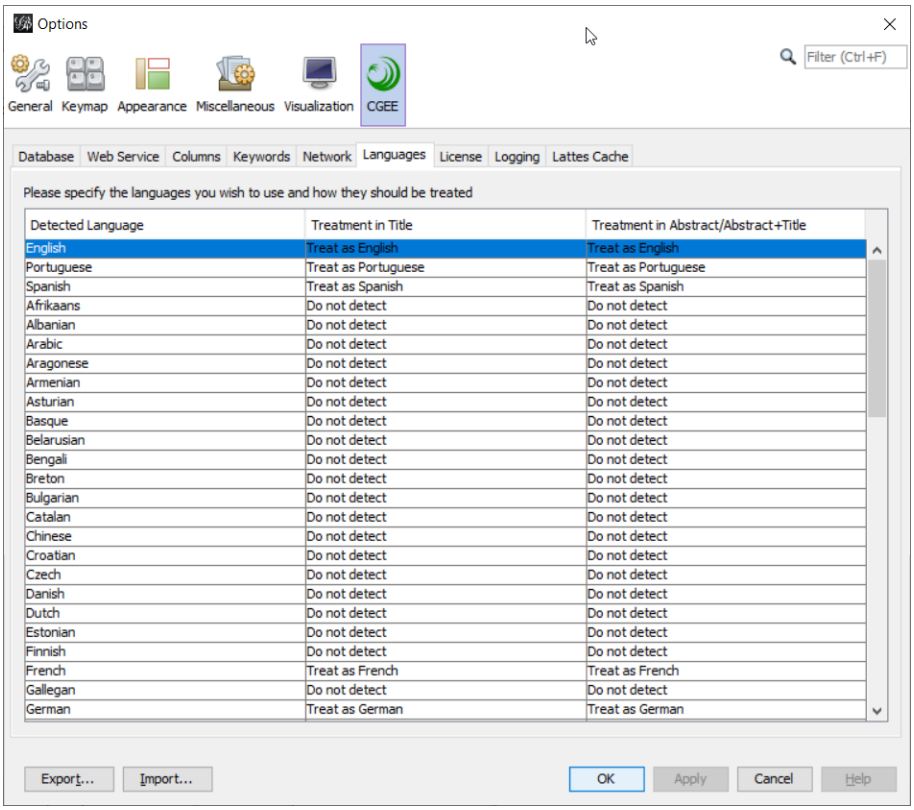


Figura 3.8 Configuração da detecção de idiomas

Deve ser observado que não existem analisadores de texto para todos os idiomas que podem ser detectados. 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*":

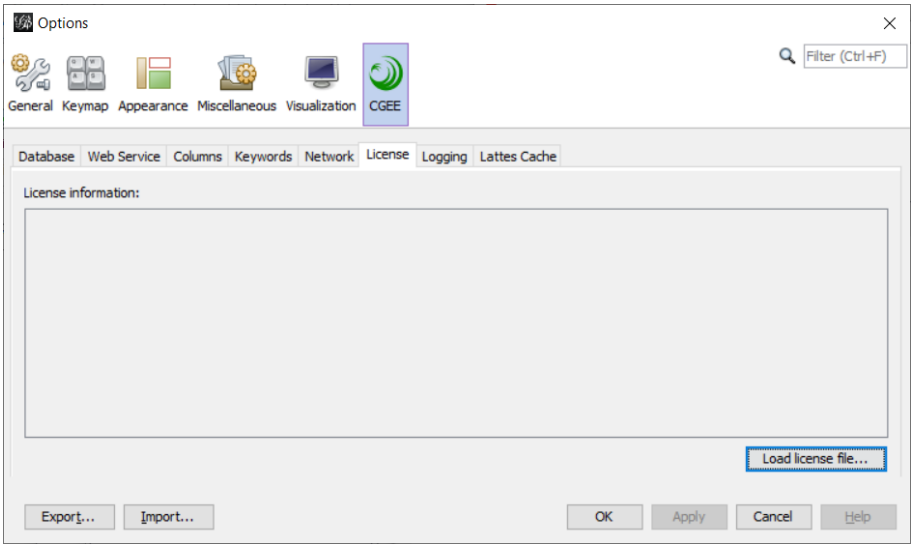


Figura 3.9 Instalação de licenças

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:

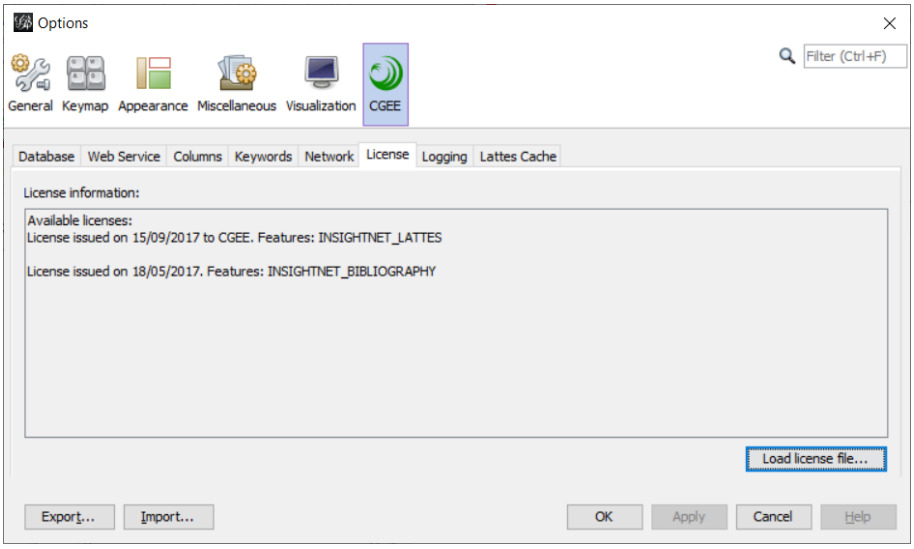


Figura 3.10 Indicação das licenças disponíveis

3.8 Protocolos de execução

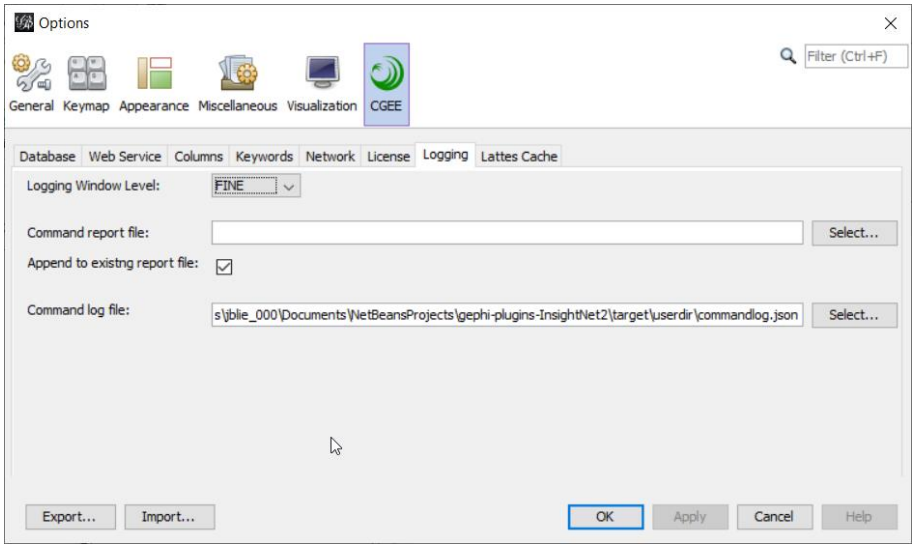


Figura 3.11 Configuração do protocolo de execução

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 configuração 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 *Command 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 quantidade de currículos, do tamanho total dos currículos e da idade máxima. Esses limtes podem ser configurados nesta aba:

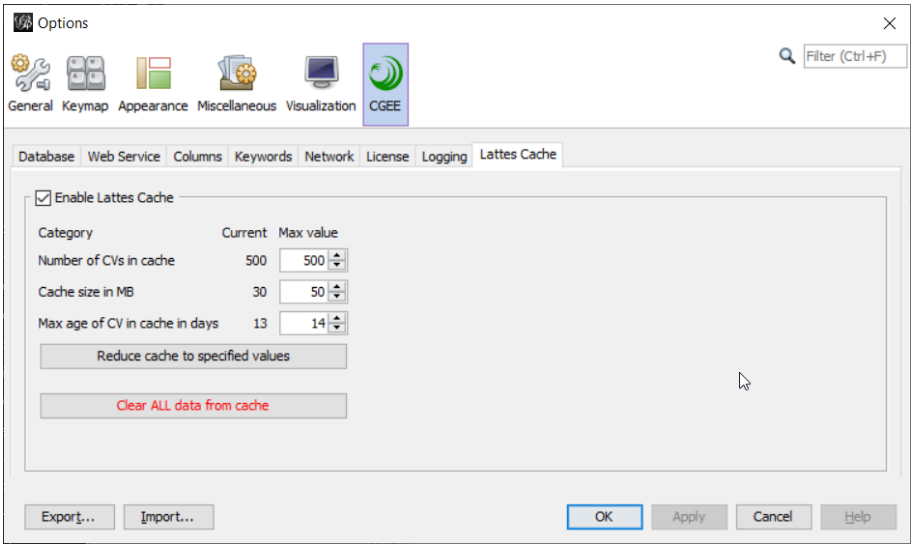


Figura 3.12 Configuração da memória cache do módulo Lattes

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).

4 Conceitos gerais do uso do CGEE Insight Net

4.1 Fluxo de trabalho

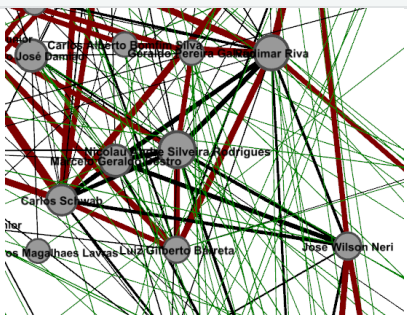
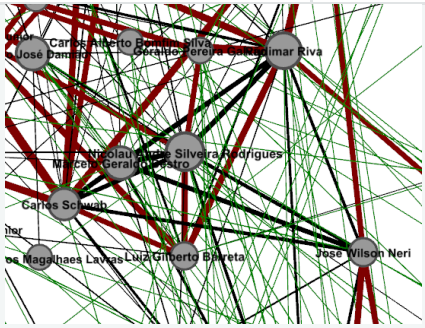
Ao utilizar o *CGEE insight Net*, deve ser considerado que ele trabalha com três repositórios de informações:

- 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*, através 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 bibliográfica, 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

Passo	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.

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 semântica 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*:

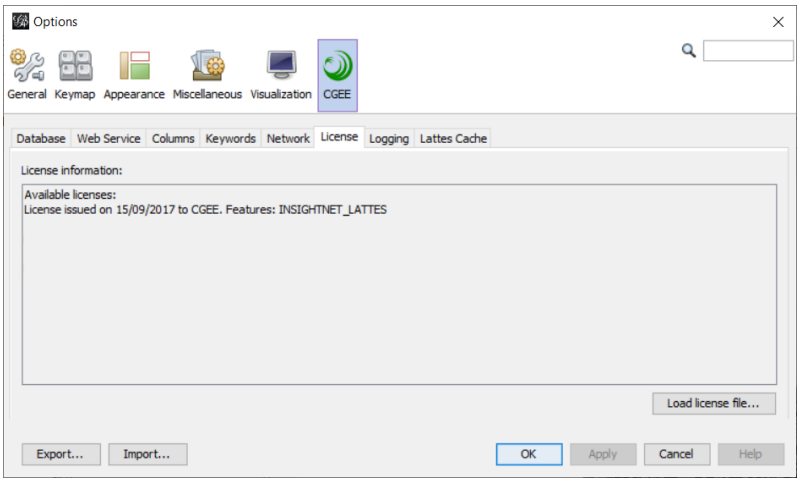


Figura 5.1 Licença requerida para o módulo de redes de Currículos Lattes

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

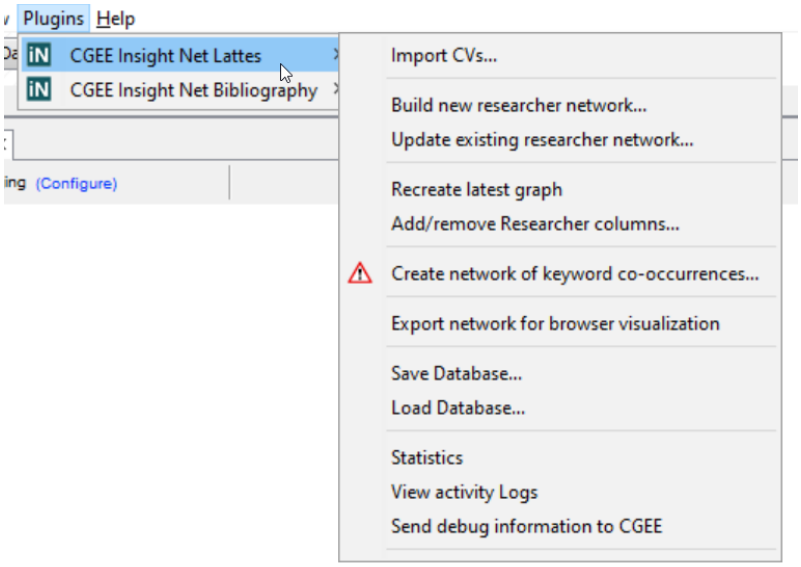


Figura 5.2 Sub-menu CGEE Insight Net Lattes

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:

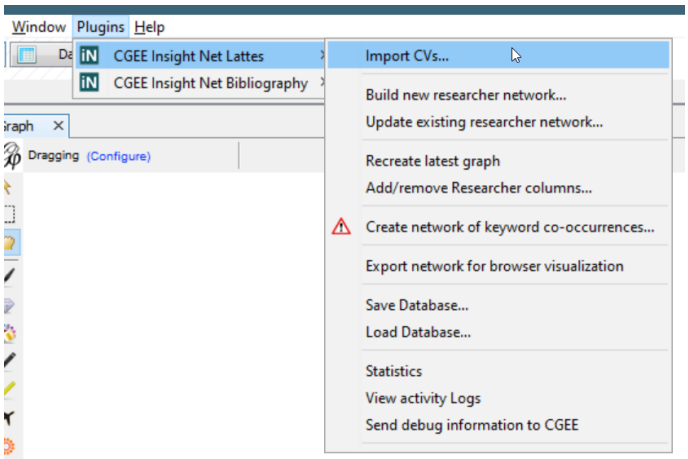


Figura 5.3 Menu de importação de Currículos

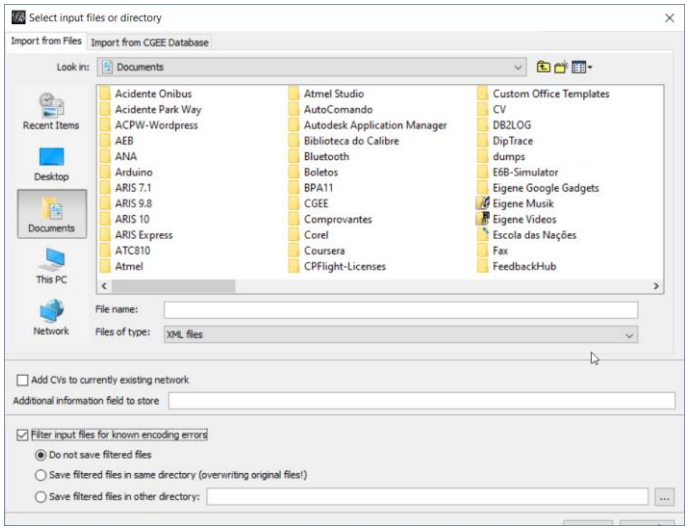


Figura 5.4 Diálog de importação de Currículos em Arquivos

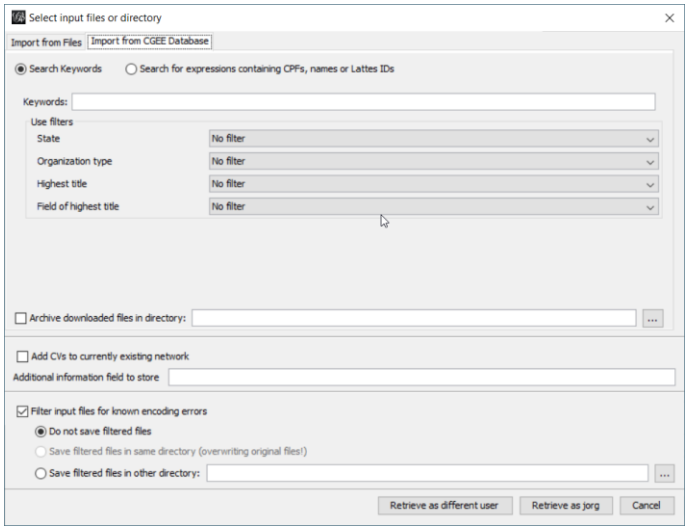


Figura 5.5 Diálogo de importação de Currículos do Banco de Dados

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 [5].

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

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 dados de currículos do CGEE. Neste caso, o diálogo oferece duas funcionalidades para recuperar currículos Lattes. Essas funcionalidades serão descritas em seguida.

5.1.2.1 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 [6]. Adicionalmente, os currículos obtidos podem ser filtrados por "Unidade da Federação", "Tipo de organização", "Maior titulação" e "Área da maior titulação":

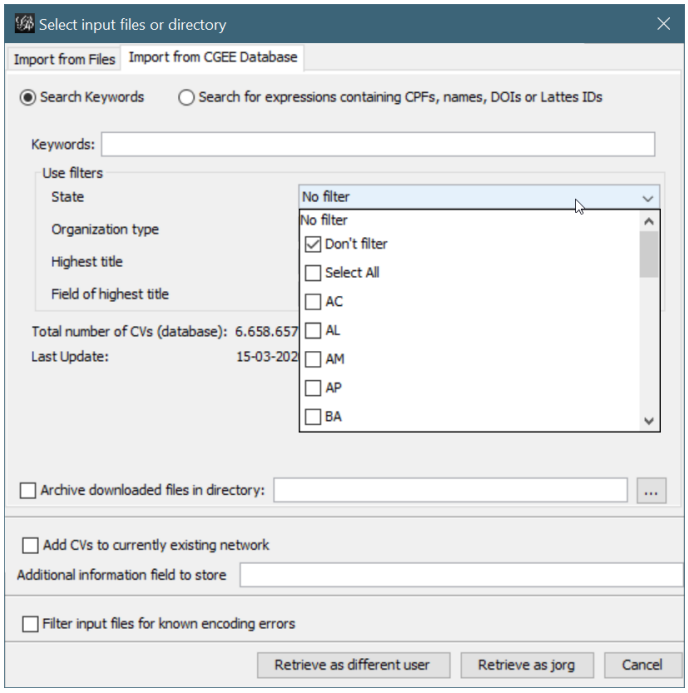


Figura 5.6 Importação de dados do Banco de dados do CGEE

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 currículos 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.

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

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.

5.1.2.2 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:

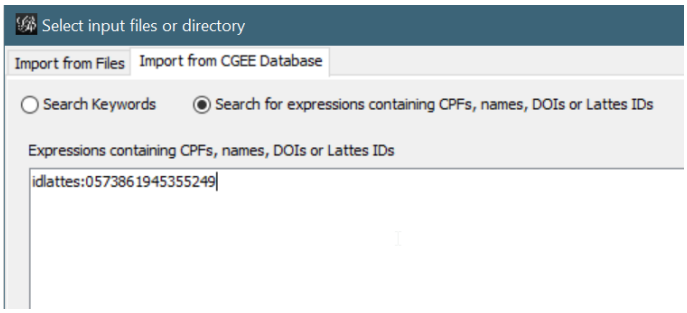


Figura 5.7 Recuperação de currículos por expressão de busca

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`

5.1.2.3 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

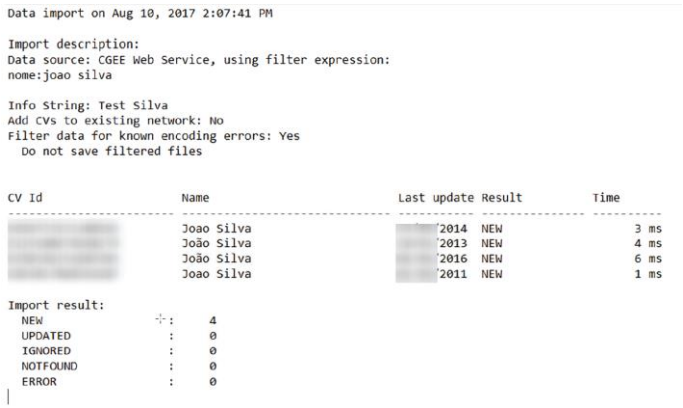


Figura 5.9 Arquivo SUMMARY.TXT, descrevendo os insumos e o resultado da importação

5.1.3 Opções comuns

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

5.1.3.1 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.

5.1.3.2 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 outros. O atributo "*info*" dos nós é preenchido com o valor especificado no campo "*Additional information field to store*" durante a importação.

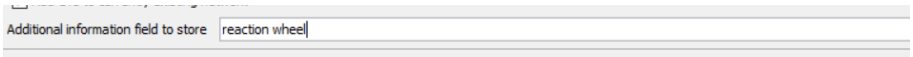


Figura 5.11 Opção do campo adicional de informação

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 currículo é importado várias vezes com valores diferentes no campo "info", esses valores serão adicionados e separados com o caractere "/".

5.1.3.3 Limpeza dos dados

Para permitir o processamento de arquivos contendo alguns tipos de erros identificados na base de currículos, 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.

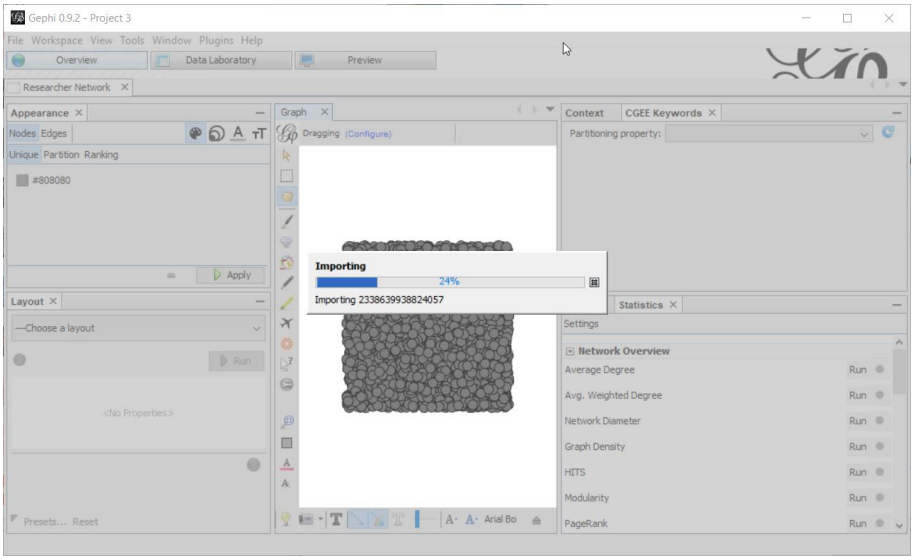


Figura 5.13 Importação dos currículos Lattes

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

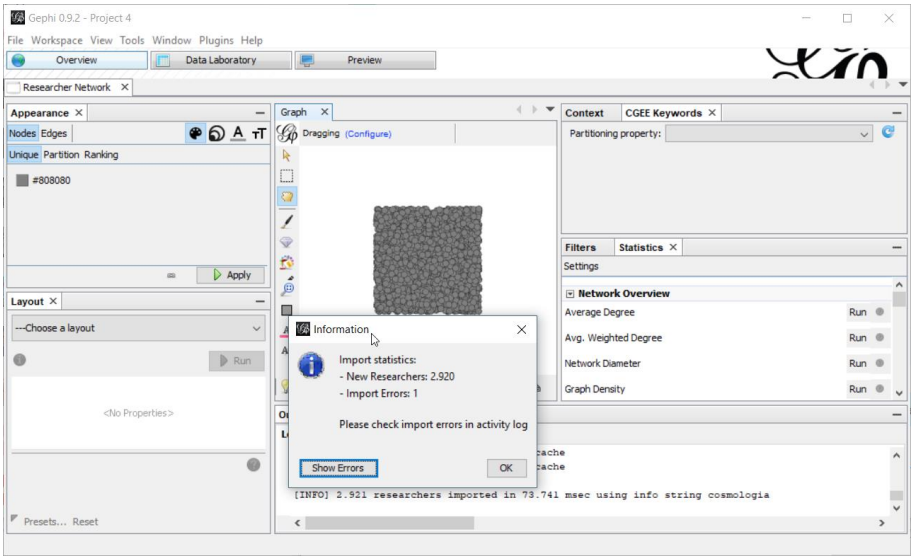


Figura 5.14 Resultado da importação dos currículos Lattes

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).

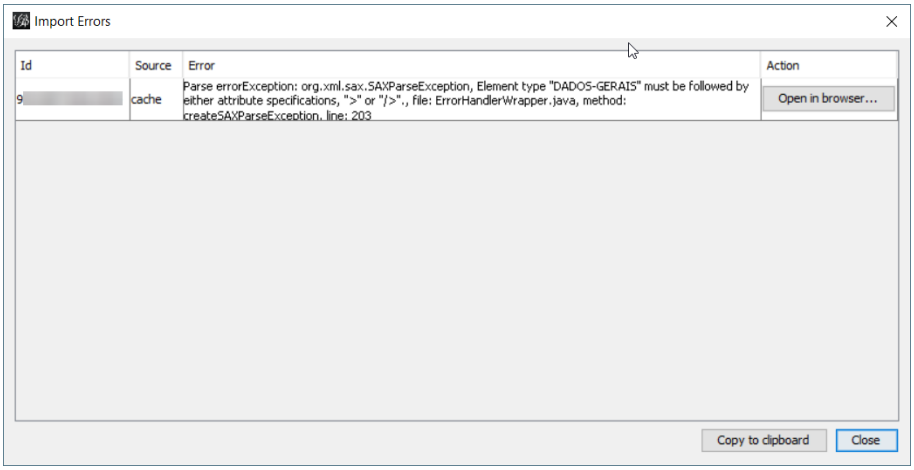


Figura 5.15 Diálogo de erros de importação

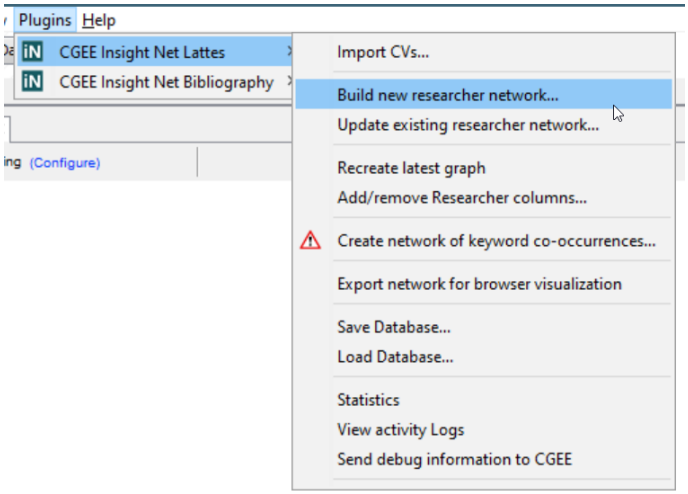
Adicionalmente, o protocolo de execução (ver seção Protocolos de execução) registra informações sobre o andamento da importação, de acordo com o grau de detalhe especificado na tela de configuração (ver seção :options-log).

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:



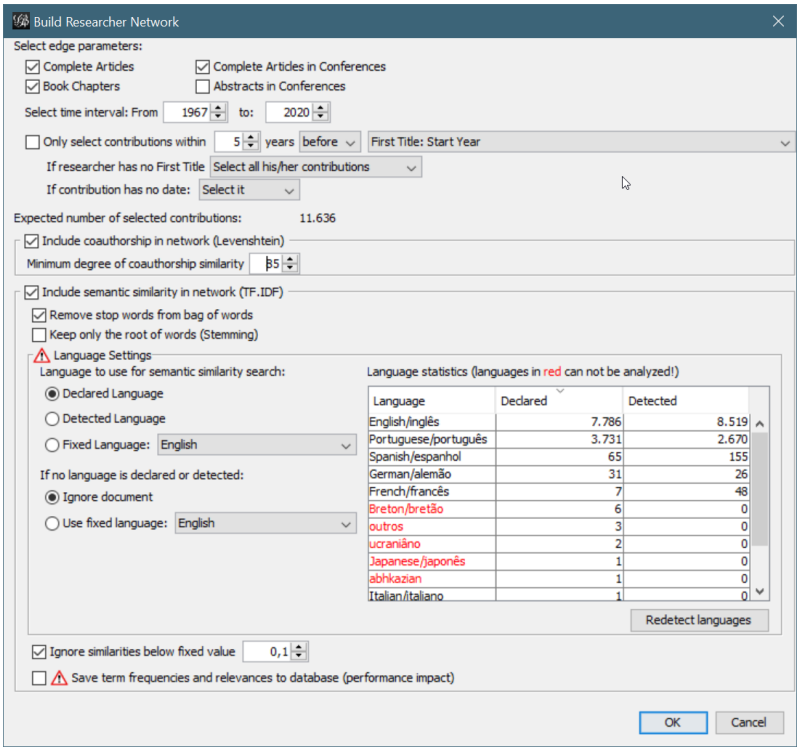


Figura 5.16 Menu e diálogo da formação da rede de currículos Lattes

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*:

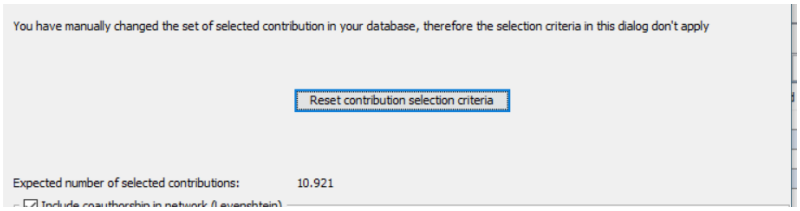


Figura 5.17 Informação sobre a alteração manual do escopo de cálculo da rede

Para desfazer as alterações manuais do escopo de rede, o usuário pode clicar no botão `*Reset contribution 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 selecionadas.

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 permanecem desabilitadas:

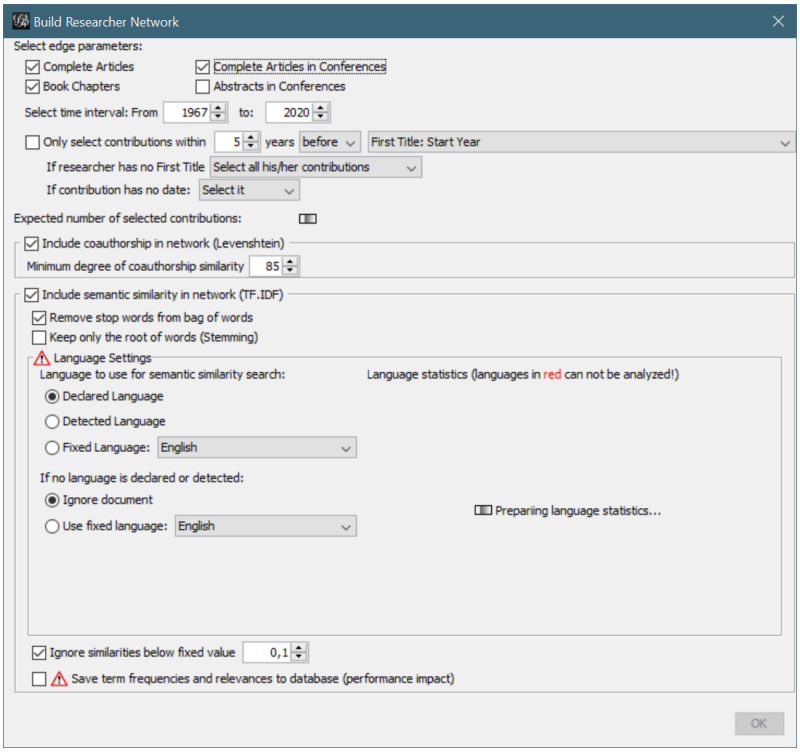


Figura 5.18 Recálculo da quantidade de contribuições selecionadas

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 coautorias.

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 similaridade 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.

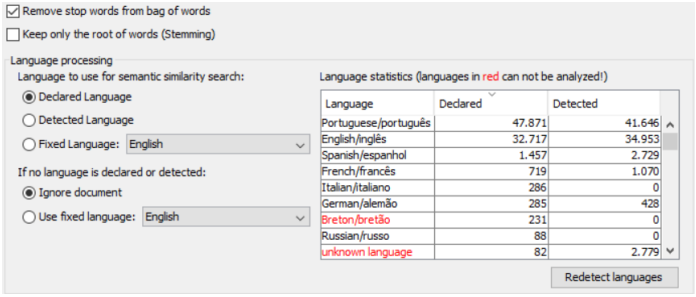


Figura 5.19 Estatística de idiomas detectados e declarados

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

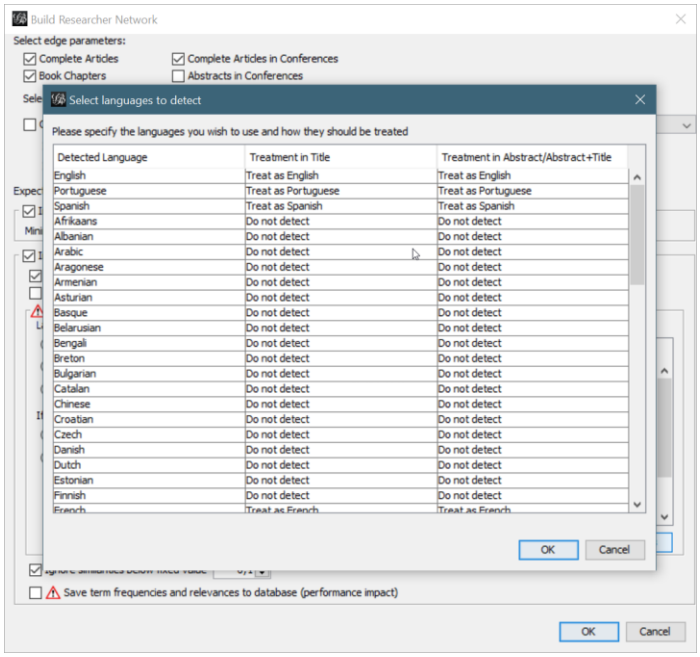


Figura 5.20 Nova detecção de idiomas

- O cálculo da similaridade semântica gera arestas entre praticamente qualquer par de pesquisadores. 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 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 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:

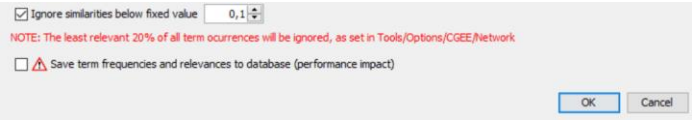


Figura 5.21 Aviso sobre configuração de limite inferior de relevância

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:

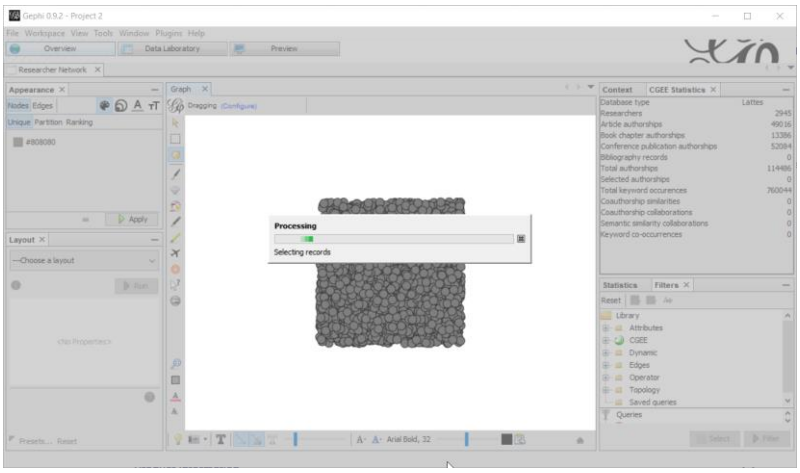


Figura 5.22 Seleção das contribuições e pré-processamento

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:

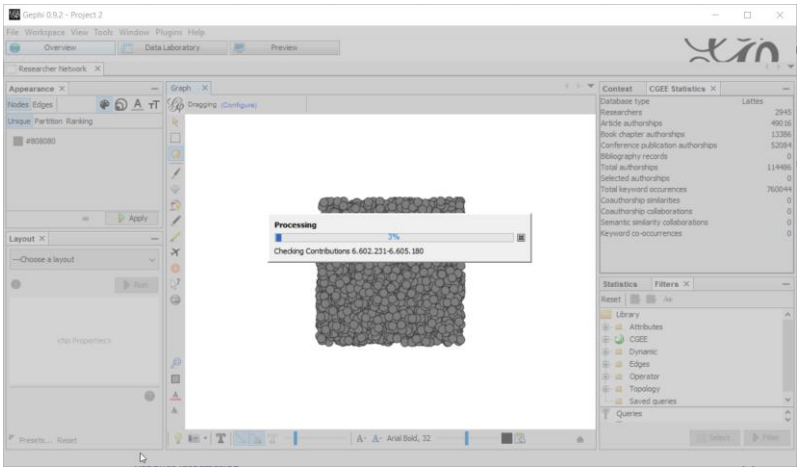


Figura 5.23 Processamento da pesquisa por similaridade

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

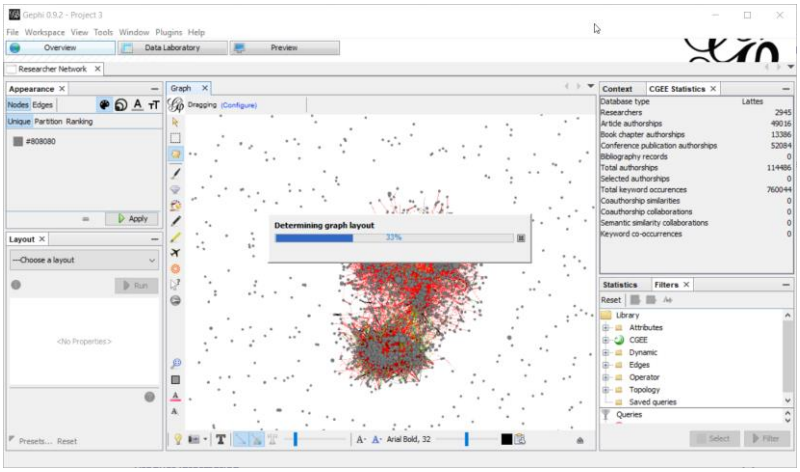


Figura 5.24 Pós-processamento e montagem da rede na tela

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…"*:

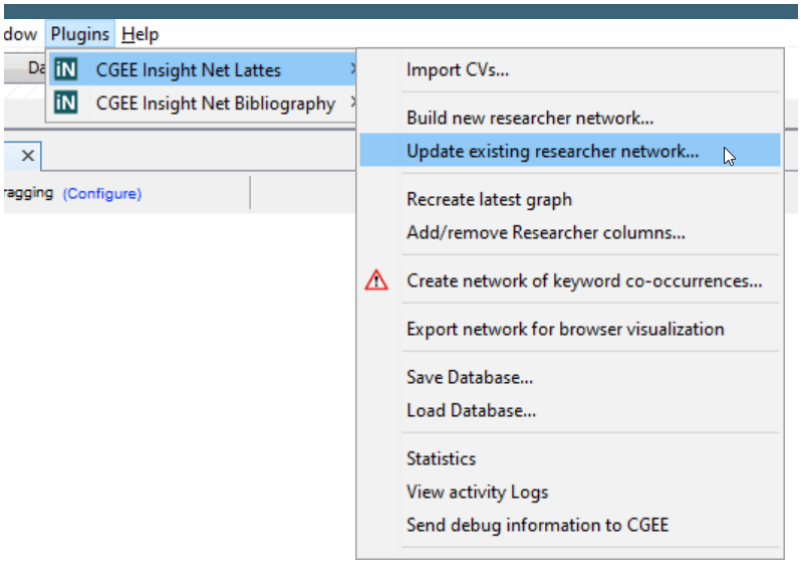


Figura 5.25 Atualização da rede já calculada

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 eventos;
- 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; - Especialização; - 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 pesquisador.

Um *termo* no sentido do parágrafo anterior corresponde, básicamente a uma palavra. 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 subconjunto que melhor atende aos seus requisitos.

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

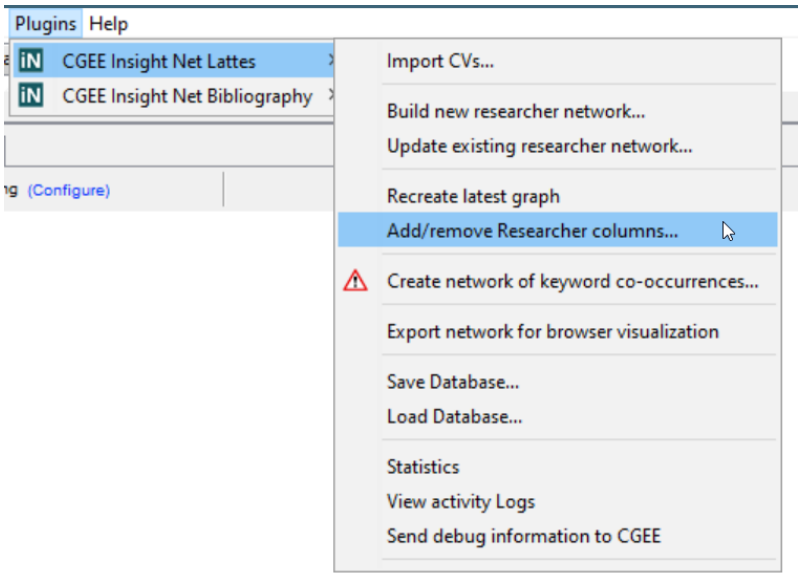


Figura 5.26 Funcionalidade para selecionar dados relevantes dos pesquisadores

Na seleção dessa funcionalidade, o seguinte diálogo é exibido:

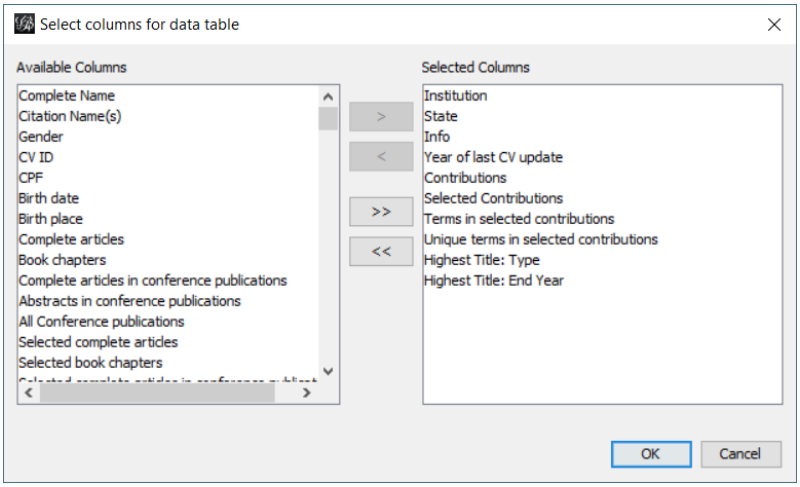


Figura 5.27 Seleção dos atributos dos pesquisadores

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:

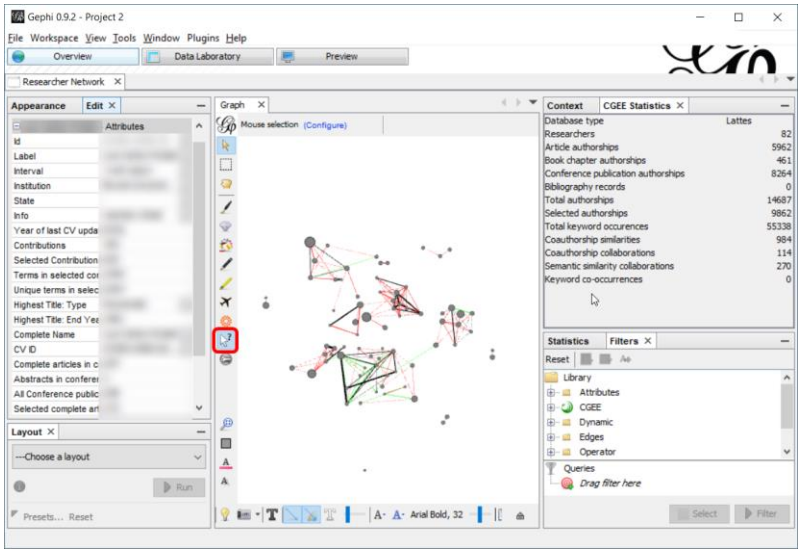


Figura 5.28 Exibição no grafo de todos os atributos habilitados do pesquisador

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  :

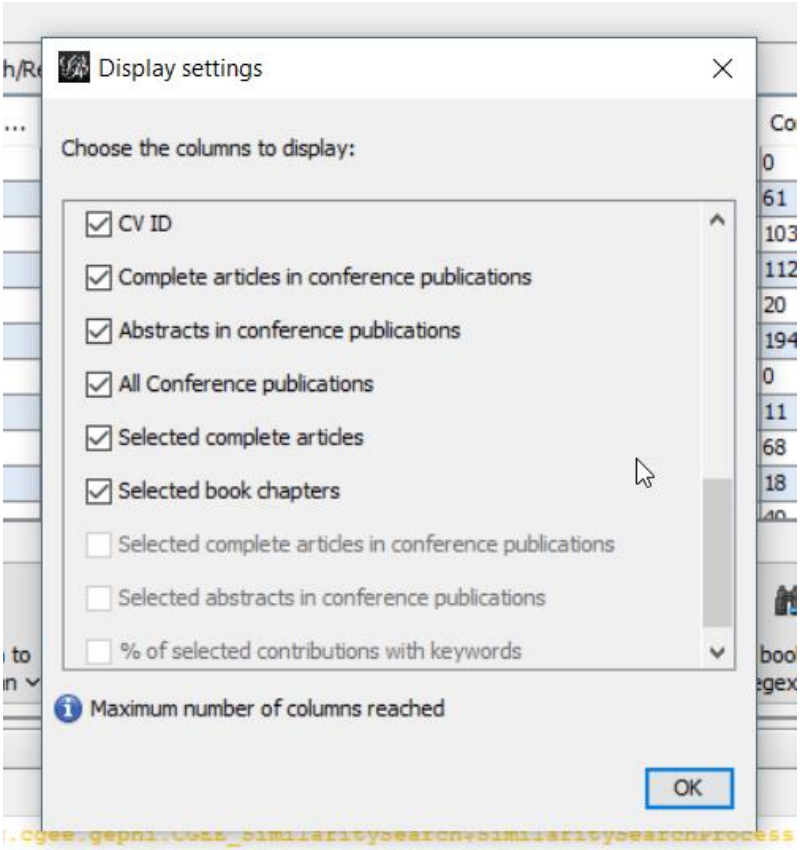


Figura 5.29 Seleção de colunas de atributos no laboratório de dados

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 contribuições são visualizados na aba do Laboratório de Dados do Gephi:

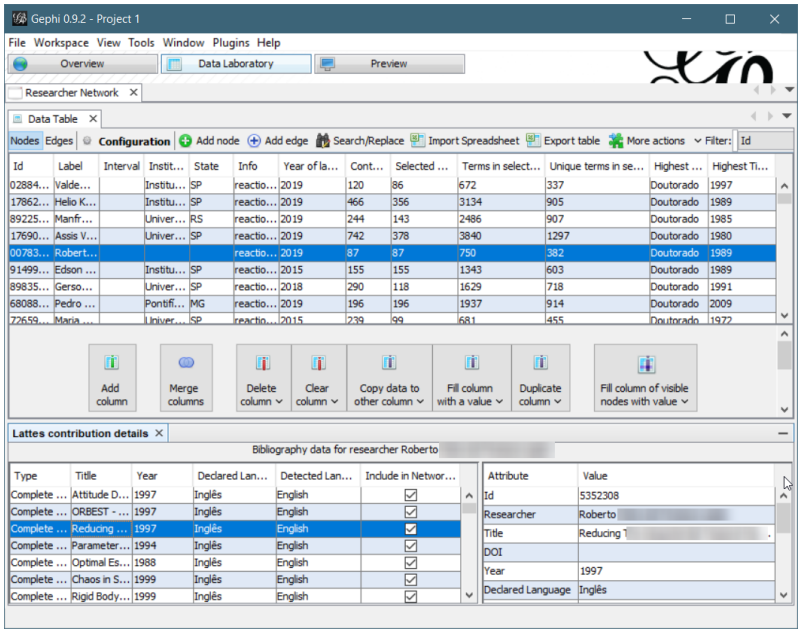


Figura 5.30 Exibição dos detalhes das contribuições bibliográfica de um Curriculo Lattes

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:

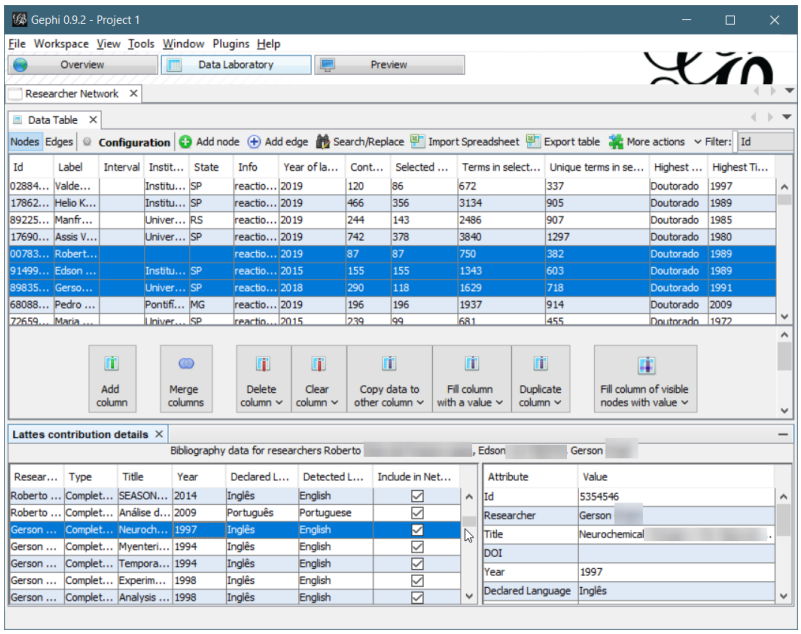


Figura 5.31 Exibição dos detalhes das contribuições bibliográfica de vários Curriculos Lattes

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 idiomas que podem ser analisados.
- Contribuições individuais podem ser incluídas ou excluidas do escopo de cálculo da

rede de coautorias 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 coautorias 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*:

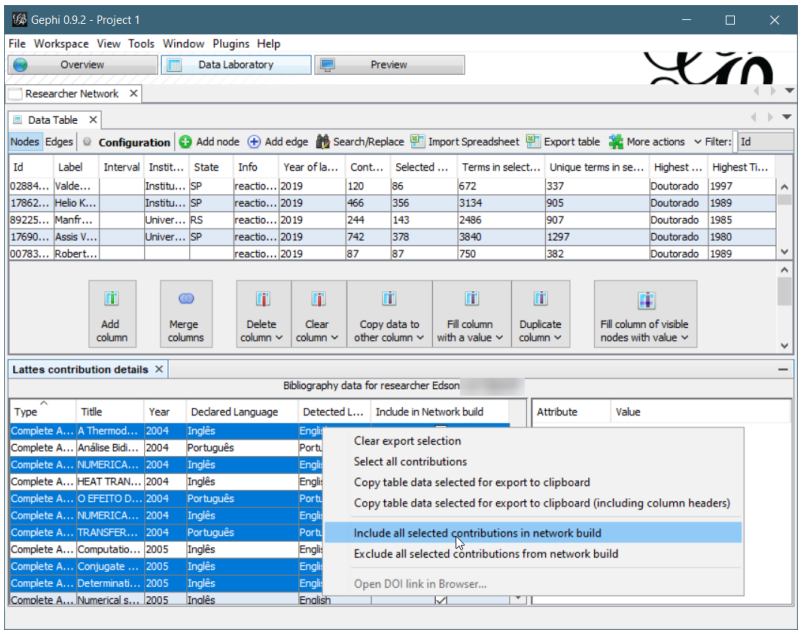


Figura 5.32 Inclusão ou exclusão de várias contribuições bibliográficas de Currículos Lattes no escopo do cálculo de rede

- 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…*

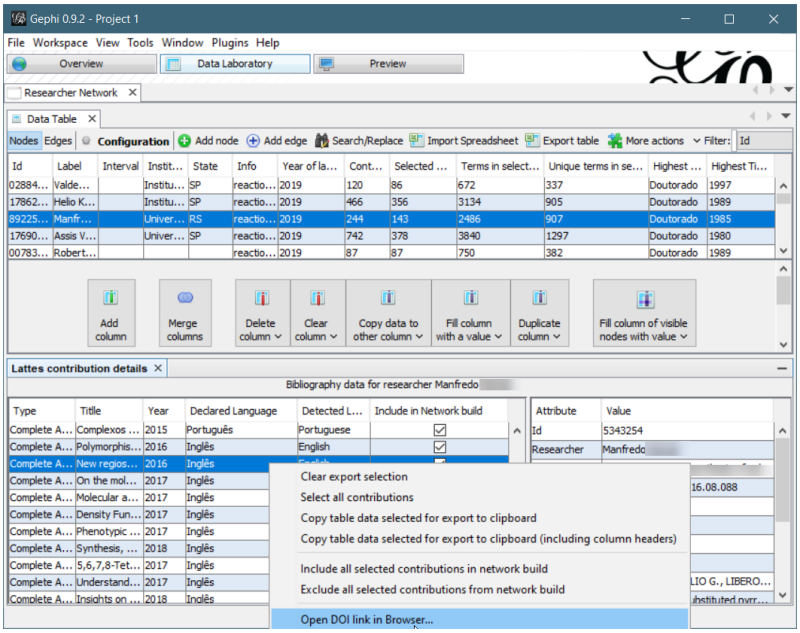


Figura 5.33 Exibir o Document Object Identifier no Browser

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 (*"abstracts"*) 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*:

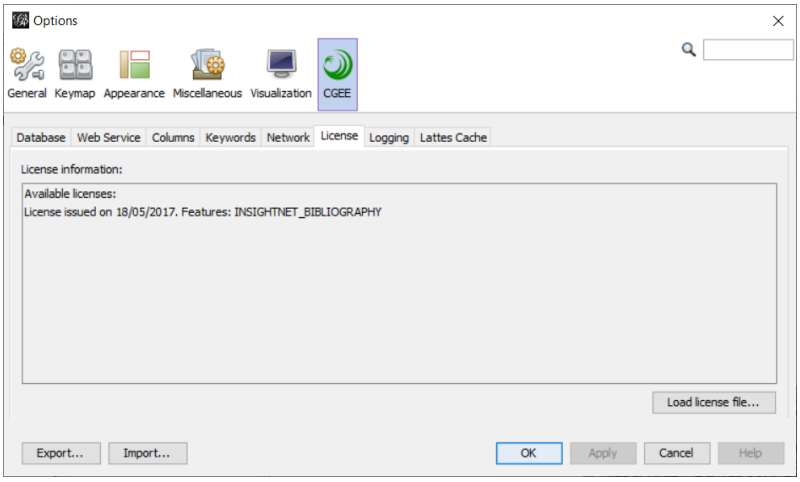


Figura 6.1 Licença requerida para o módulo de redes bibliográficas

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

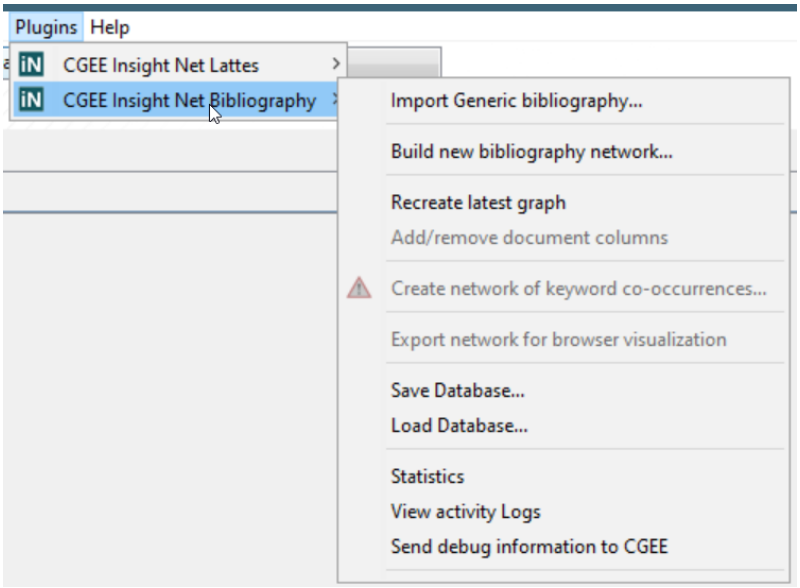
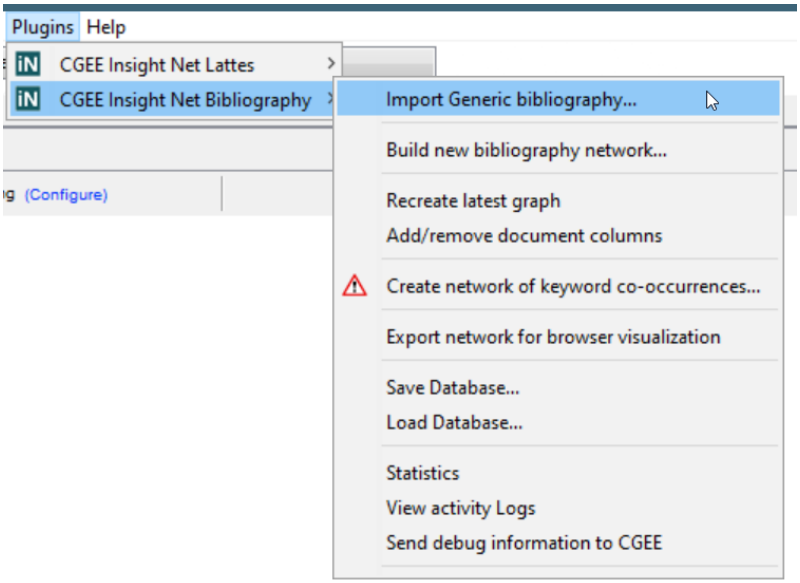


Figura 6.2 Sub-menu CGEE Insight Net Bibliography

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.



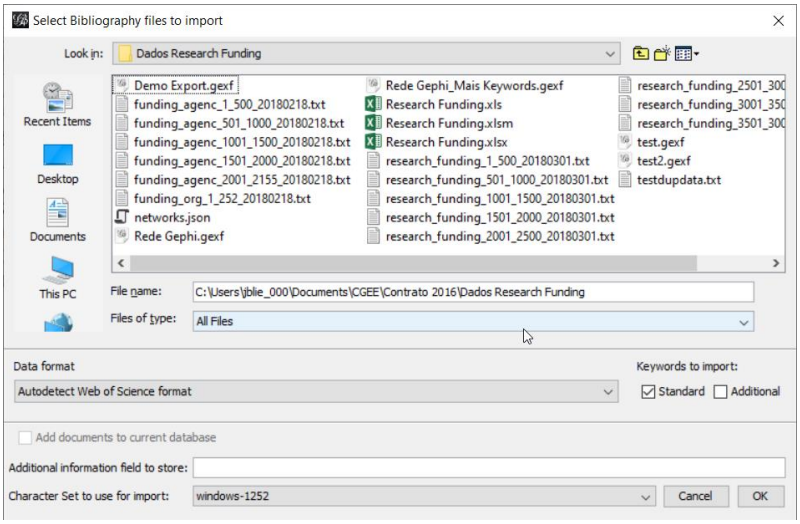


Figura 6.3 Importação de arquivos de bibliografia

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.

6.1.1.1 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 mencionadas acima, bem como a possibilidade de detectar o formato automaticamente, o que, geralmente, é a forma mais indicada:

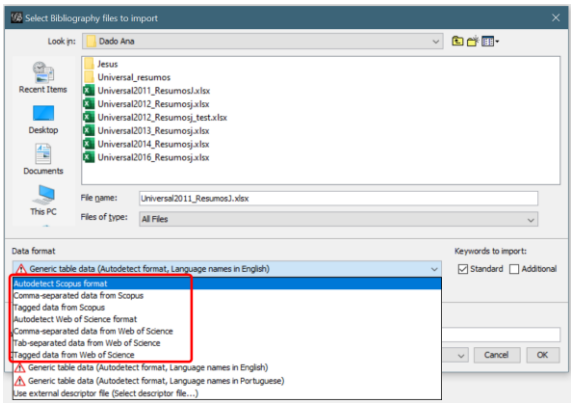


Figura 6.4 Formatos de arquivos de bibliografia dos serviços Web of Science®

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`):

6.1.1.2 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:

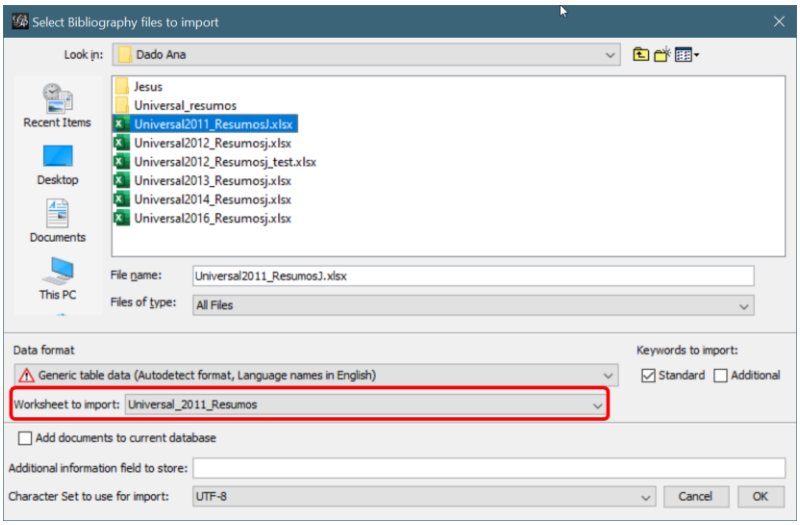


Figura 6.5 Seleção da aba da planilha excel no formato "Generic Table data (Autodetect format)"

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 importação, 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 especificado.
- As colunas adicionais, que possuem um nome que não consta na tabela em seguida serão importadas 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

6.1.1.3 Formato "*Use external descriptor file*"

Para importar dados em formatos que não estão cobertos por uma das opções anteriores, o *CGEE Insight Net* oferece com esta opção a possibilidade de especificar um formato em um arquivo externo. A especificação deste arquivo consta na Seção 9.1.

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:

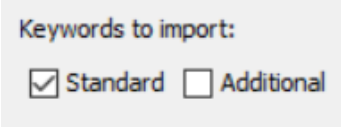


Figura 6.6 Seleção de palavras-chave a serem importadas

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 importação incremental.

Figura 6.7 Opção de importação inicial ou incremental

Se essa opção for selecionada, as referências bibliográficas importadas serão acrescentadas às informações 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 publicação (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.

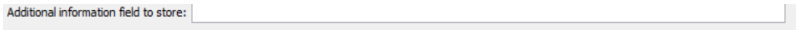


Figura 6.8 Campo adicional de informação

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.

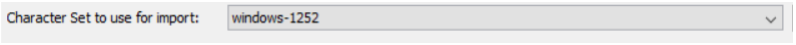


Figura 6.9 Definição do conjunto de caracteres

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).

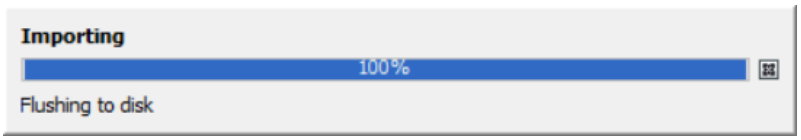


Figura 6.10 Importação de referências bibliográficas

6.2 Formação da rede

Depois da importação das referências bibliográficas na base de dados, a rede é formada a partir das pesquisas 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:

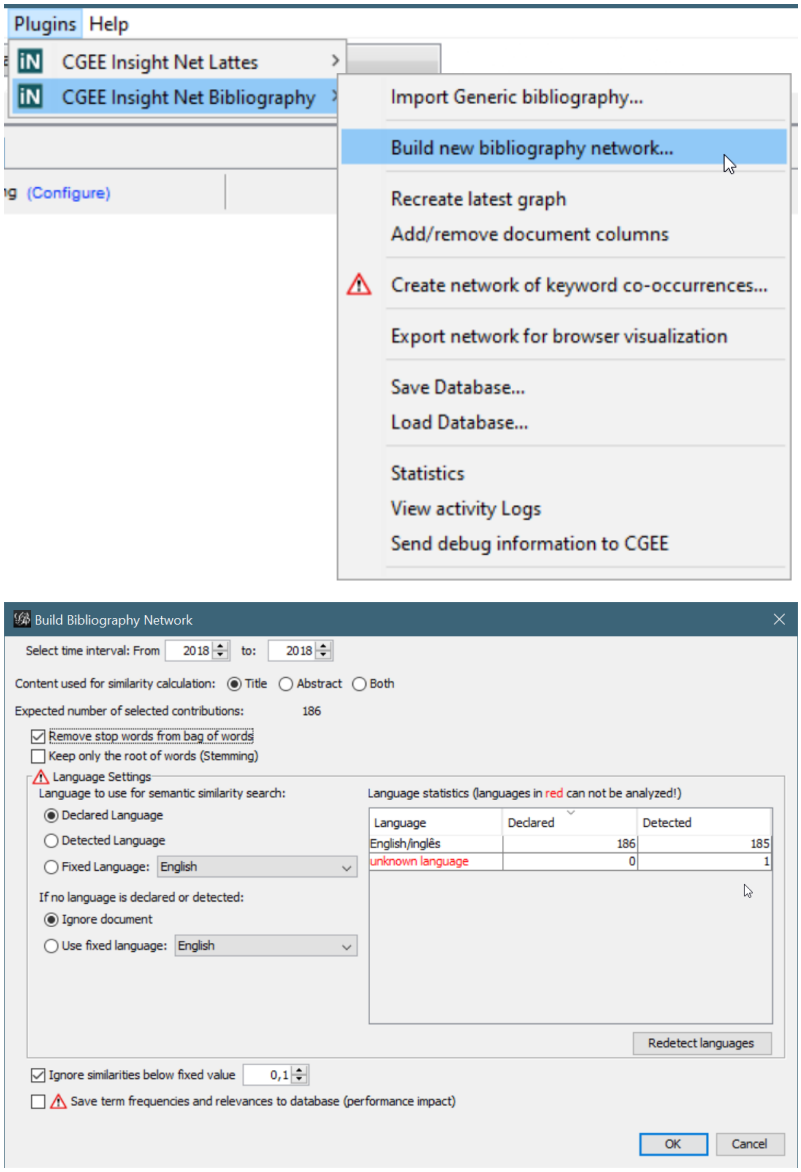


Figura 6.11 Menu e diálogo da formação da rede de referências bibliográficas

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 semântica considera apenas o título da referência bibliográfica ("*Title*"),

apenas o resumo (“*Abstract*”) 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*”.

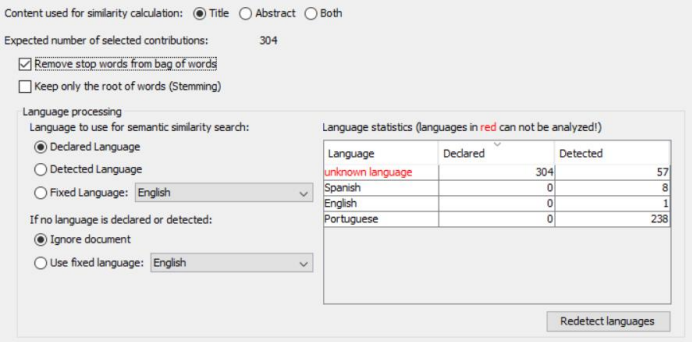


Figura 6.12 Estatística de idiomas detectados e declarados

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

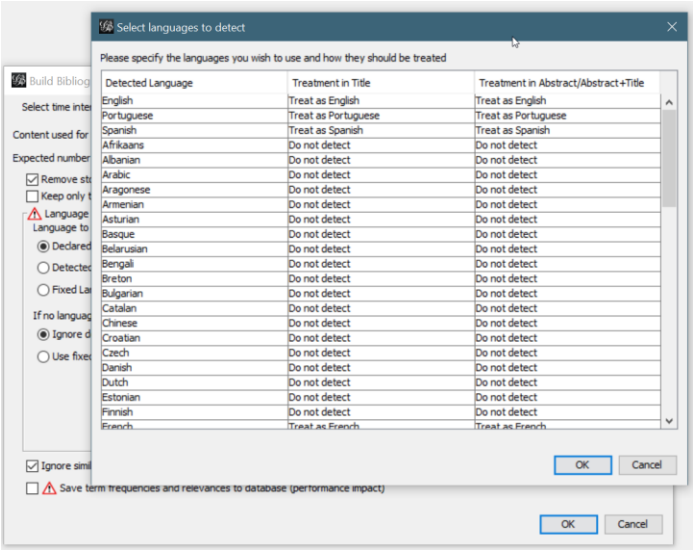


Figura 6.13 Nova detecção de idiomas

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

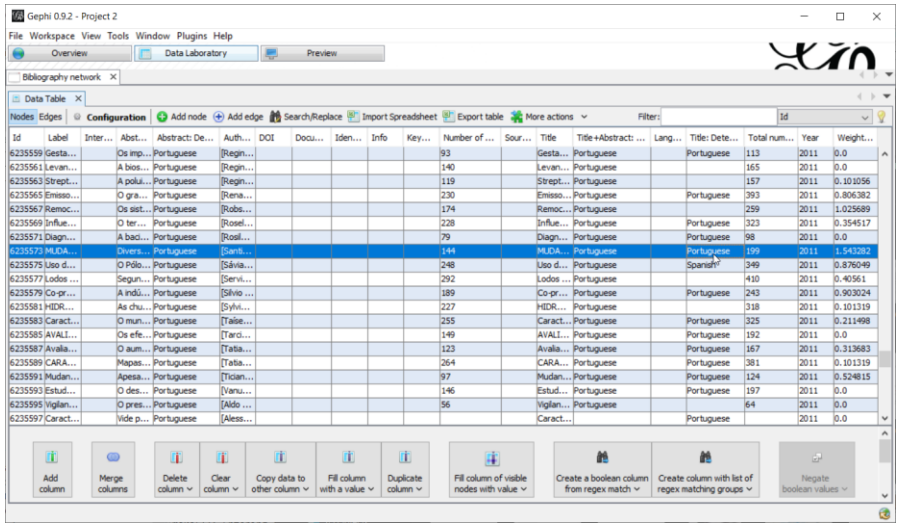


Figura 6.14 Correção manual dos idiomas das contribuições no laboratório de dados do Gephi

- 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 relevantes 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 utilizado 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 correspondente é exibida:

NOTE: The least relevant 20% of all term ocurrences will be ignored, as set in Tools/Options/CGEE/Network

Figura 6.15 Aviso sobre configuração de limite inferior de relevância

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":

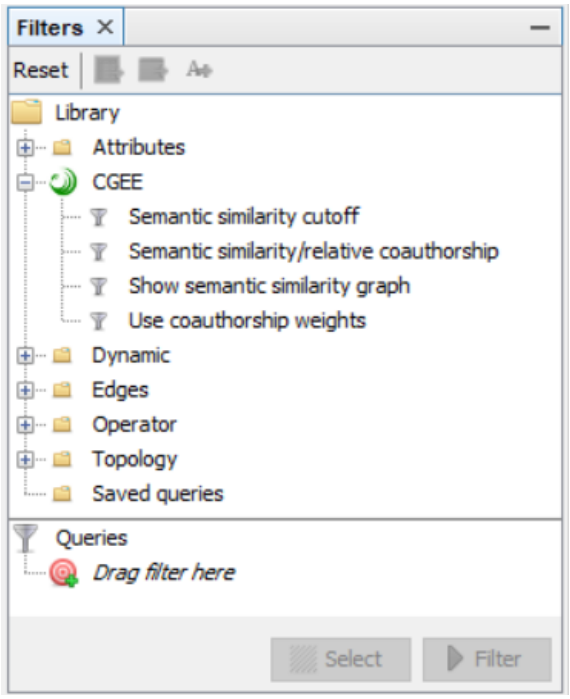


Figura 7.1 Filtros da categoria "CGEE"

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:

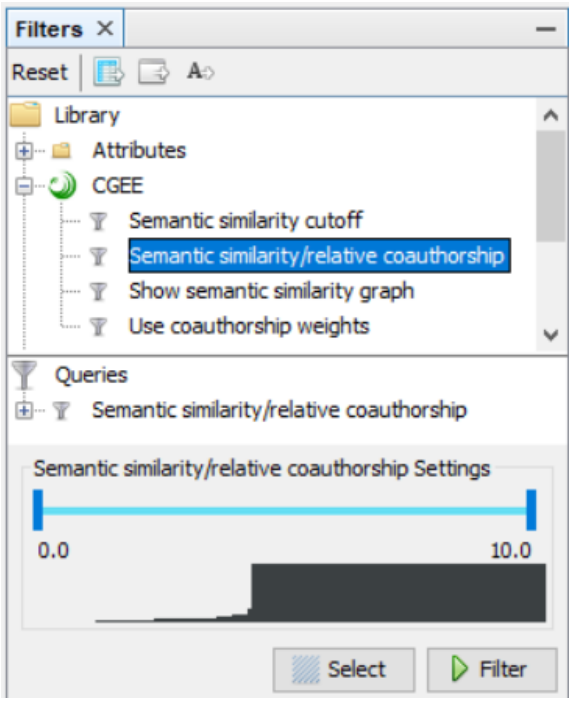


Figura 7.2 Parâmetros do filtro Semantic similarity/relative coauthorship

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.

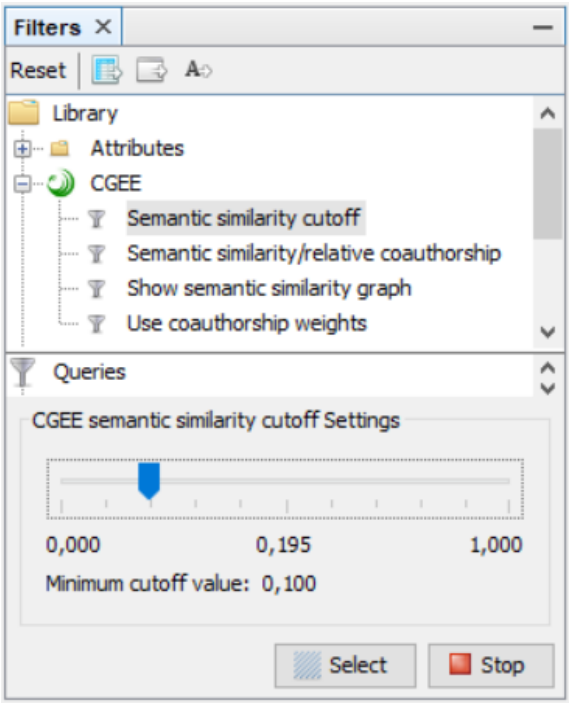


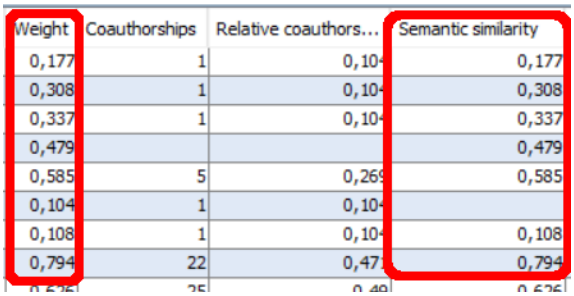
Figura 7.3 Parâmetro do filtro Semantic similarity cutoff

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.

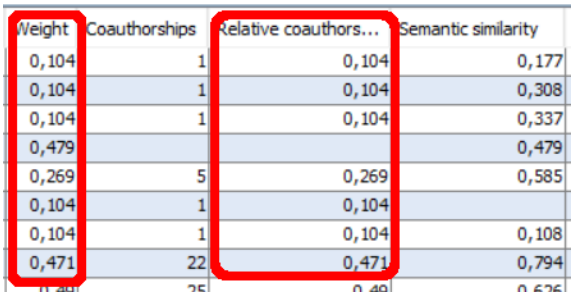
7.1.4 Filtro "*Use coauthorship weights*"

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	0,269	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*".

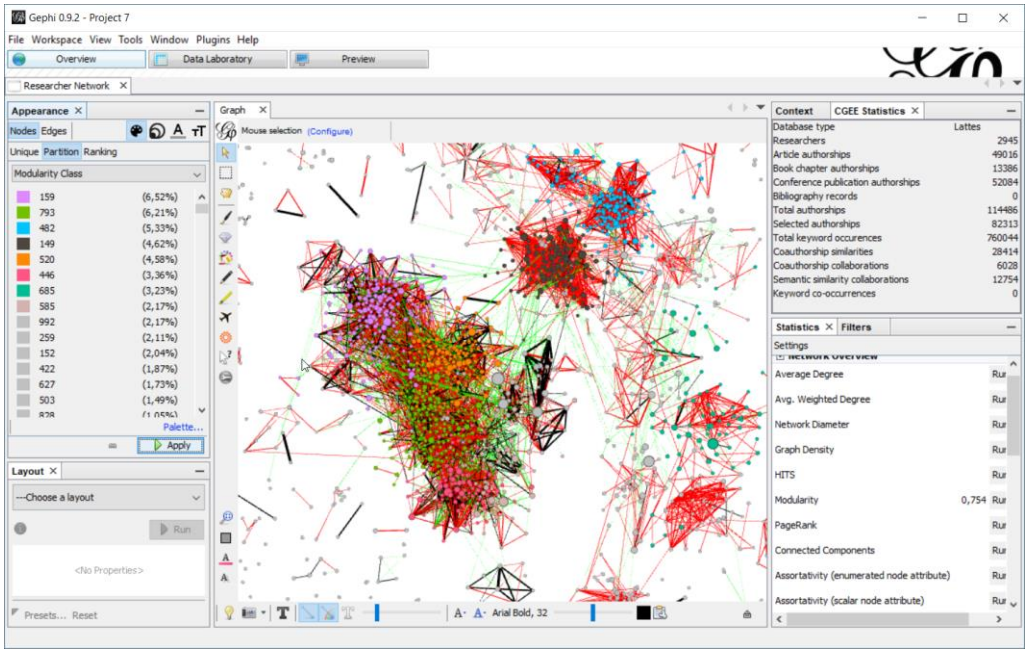


Figura 7.6 Clusters na rede gerada

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].

[Assort]

M. Newman, “*7.13 HOMOPHILY AND ASSORTATIVE MIXING*,” em *Networks. An Introduction*, New York, Oxford University Press, 2010.

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 assortatividade 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*”:

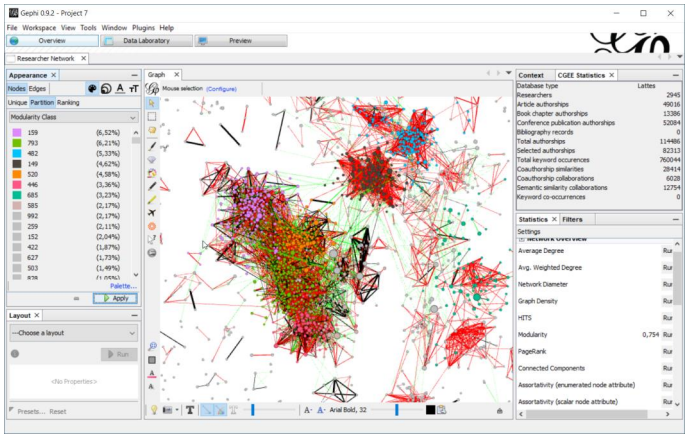
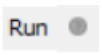


Figura 7.7 Estatísticas de Assortatividade

Dependendo do tipo de atributo, uma das duas estatísticas deve ser escolhida

- Para atributos enumerados (ou categóricos) deve ser escolhida a estatística "*Assortativity (enumerated 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 "*Assortativity (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 escolher o(s) atributo(s) com relação ao qual(is) a assortatividade deve ser calculada:

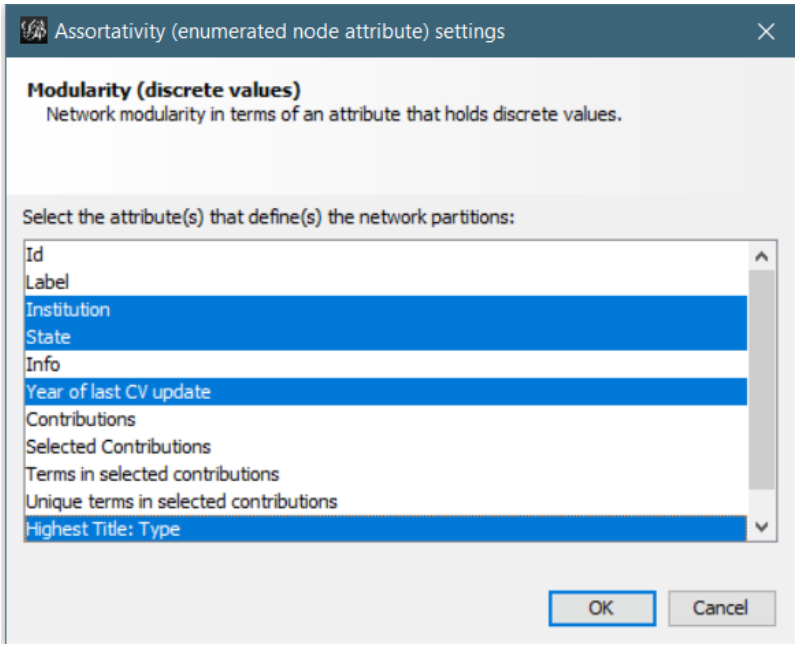


Figura 7.8 Seleção do(s) atributo(s) de assortatividade

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

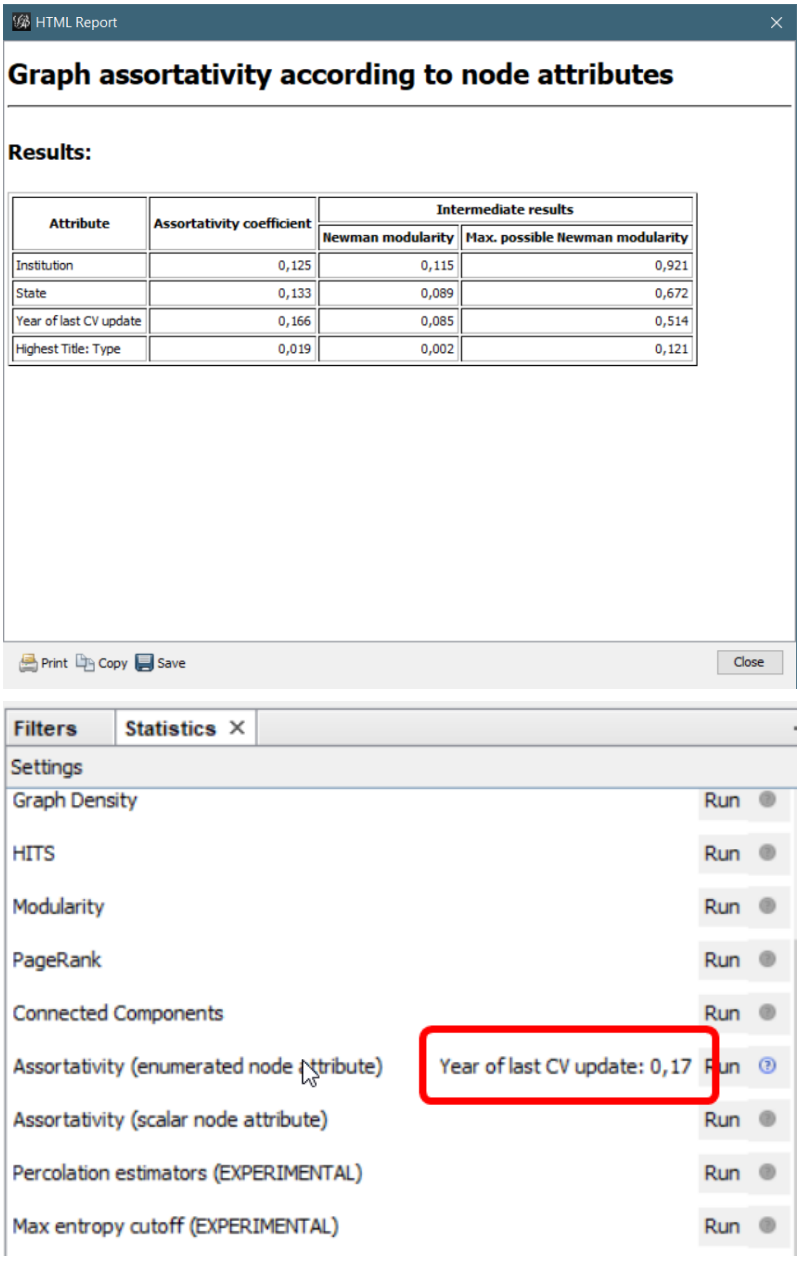


Figura 7.9 Resultados do cálculo de assortatividade

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].

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

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 extraídas 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:

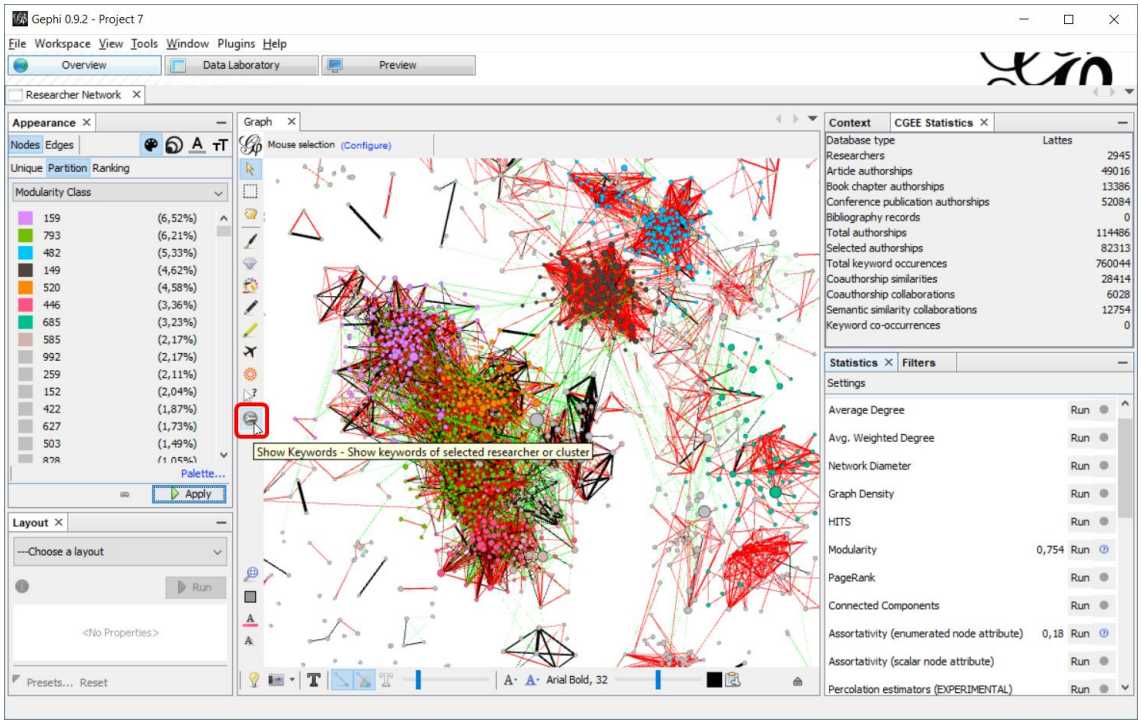


Figura 7.10 Seleção da funcionalidade "Palavras-chave"

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:

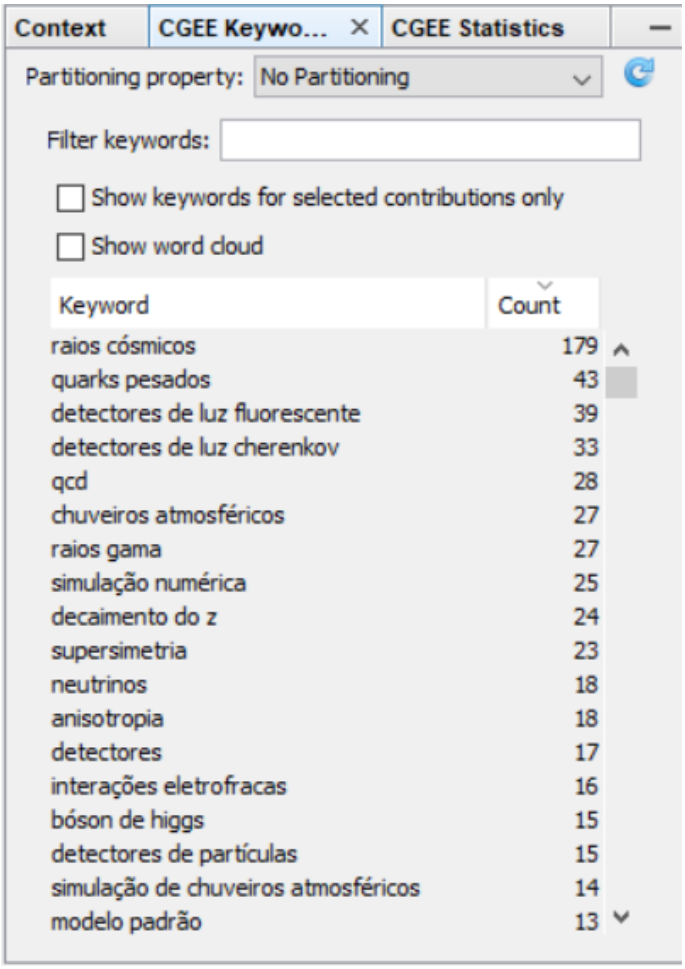


Figura 7.11 Janela de palavras-chave com frequências

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.

7.4.1.1 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:

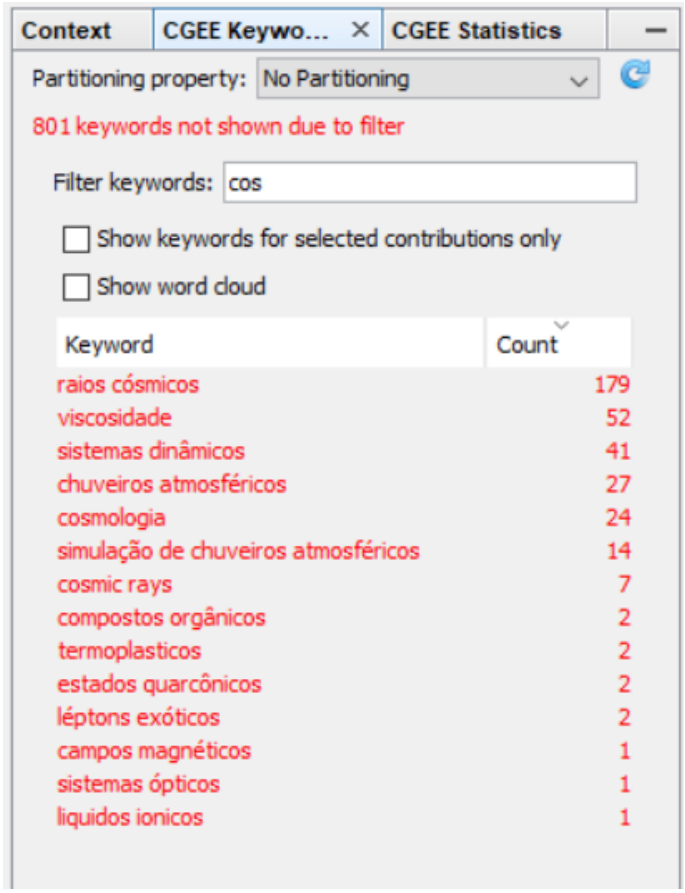


Figura 7.12 Filtragem de palavras-chave por termo

7.4.1.2 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 contextual 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:

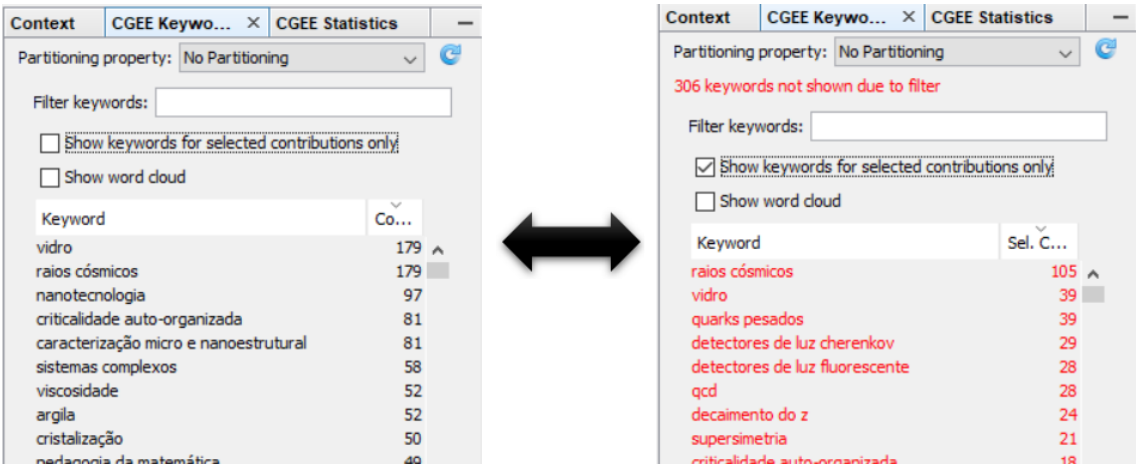


Figura 7.13 Filtragem das palavras-chave pelas contribuições selecionadas

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.

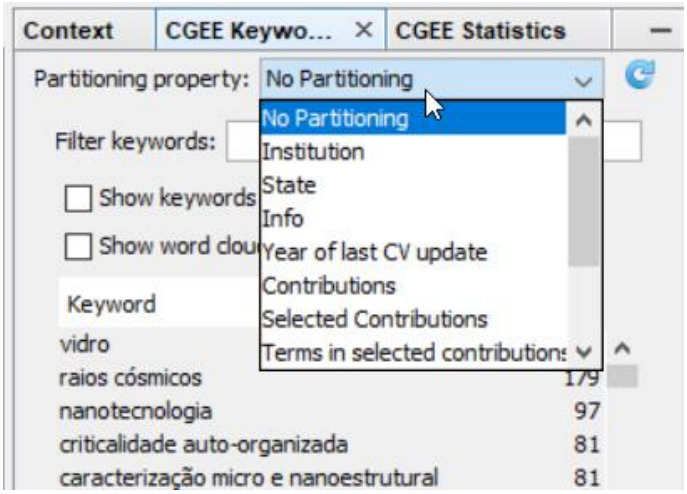


Figura 7.14 Seleção do atributo que define a partição dos nós

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 particionamento 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:

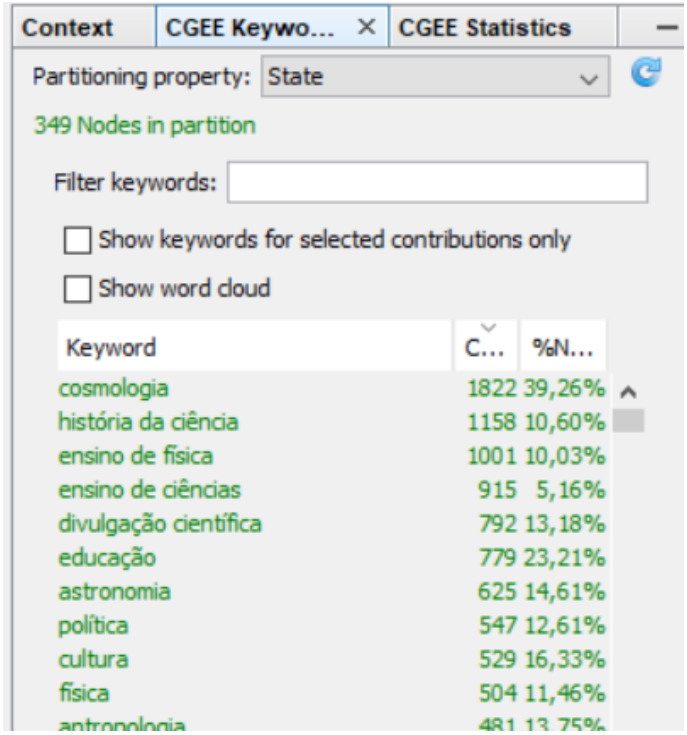


Figura 7.15 Palavras-chave de um cluster de 349 nós

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 7.4.1.1 exibe as palavras-chave na cor laranja. Observe-se que no exemplo em baixo a lista foi configurada para mostrar a quantidade absoluta de nós (ver Seção 3.4):

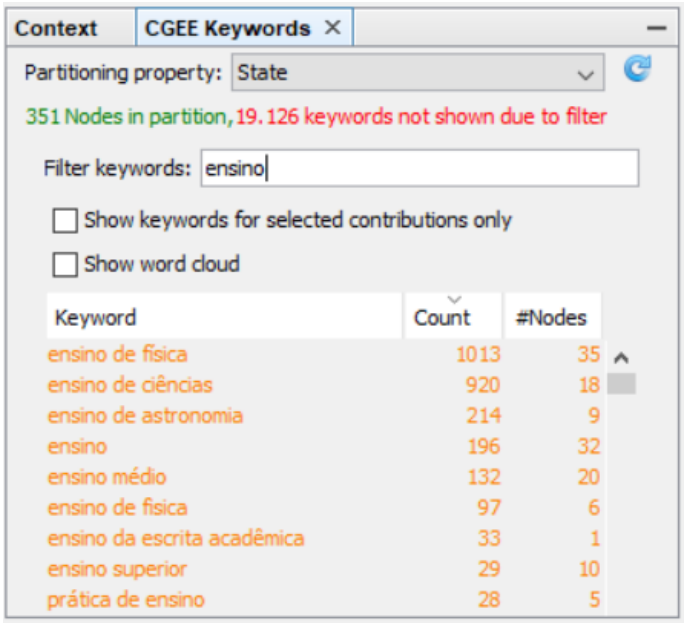


Figura 7.16 Lista de palavras-chave por partição, filtrada

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:

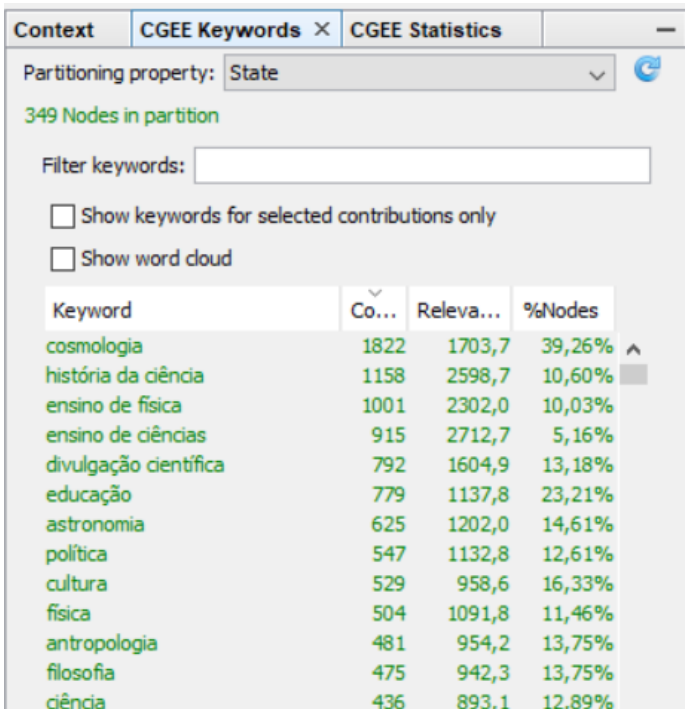


Figura 7.17 Relevância das palavras-chave exibidas na lista

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:

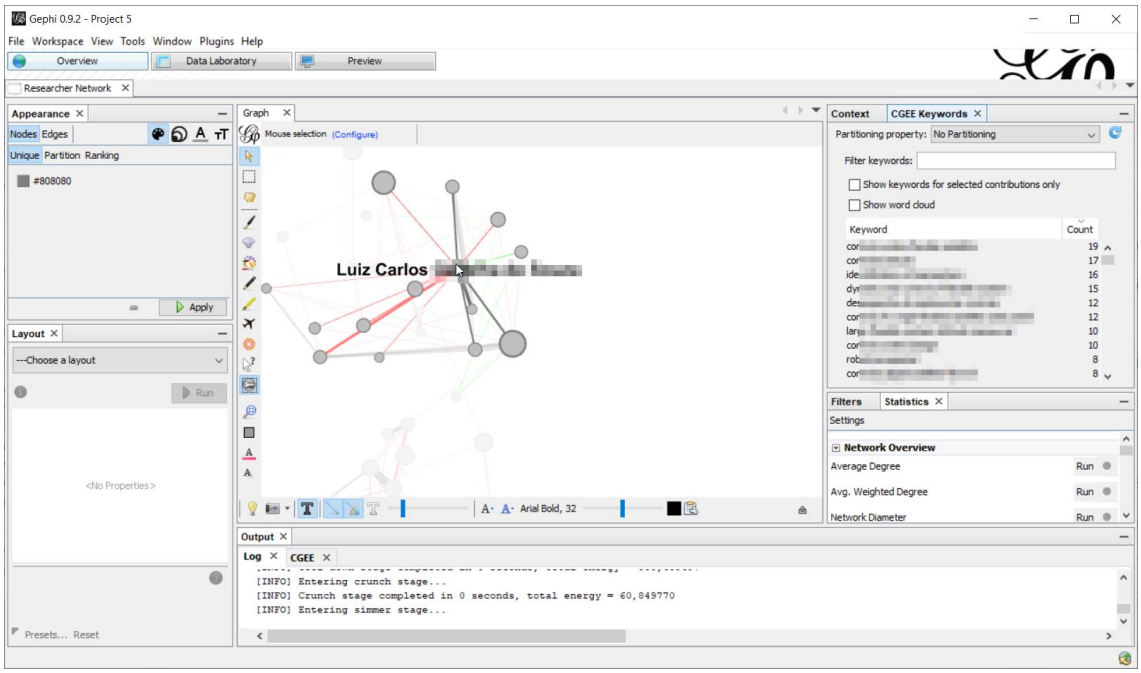


Figura 7.18 Exibição das palavras-chave do nó selecionado

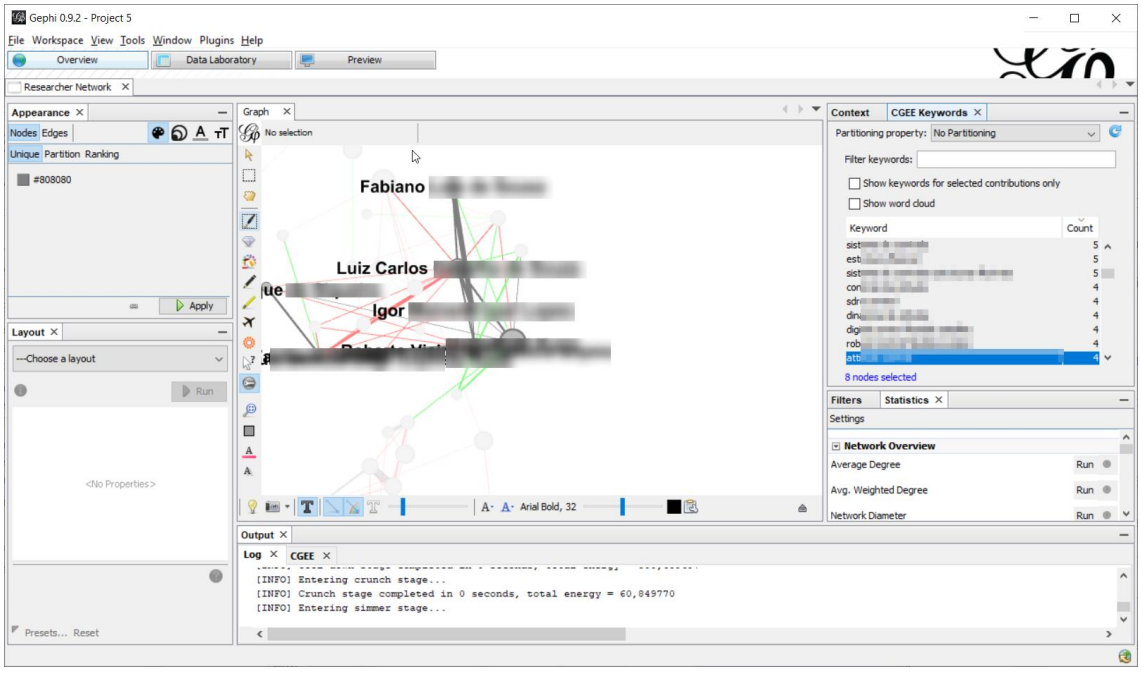


Figura 7.19 Exibição dos nós que usam a palavra-chave selecionada

A lista permite a seleção de uma única palavra-chave com um clique do botão esquerdo do mouse. Segurando 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 independentemente 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 adicionais:

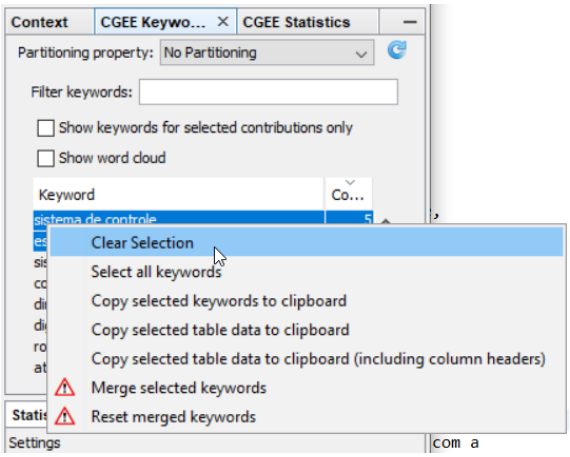


Figura 7.20 Funcionalidades adicionais na lista de palavras-chave

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:

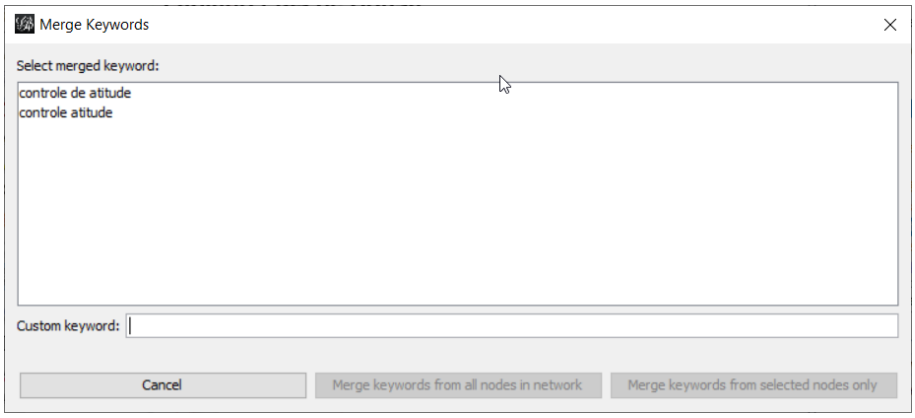


Figura 7.21 Diálogo de junção de palavras-chave

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

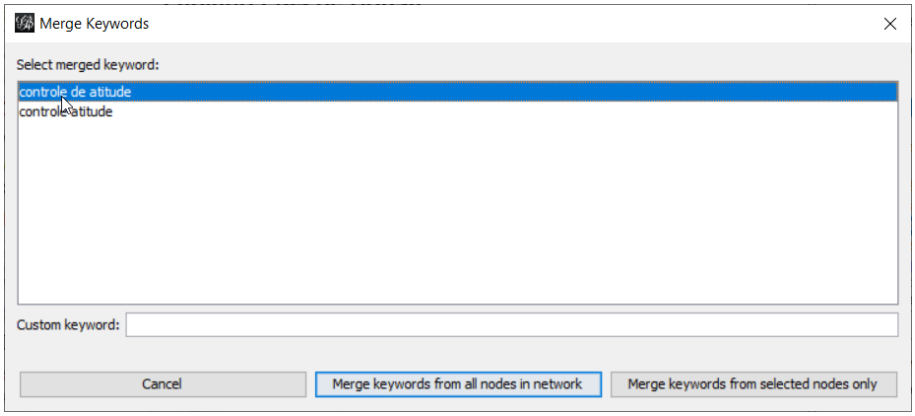


Figura 7.22 Seleção da palavra-chave juntada

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

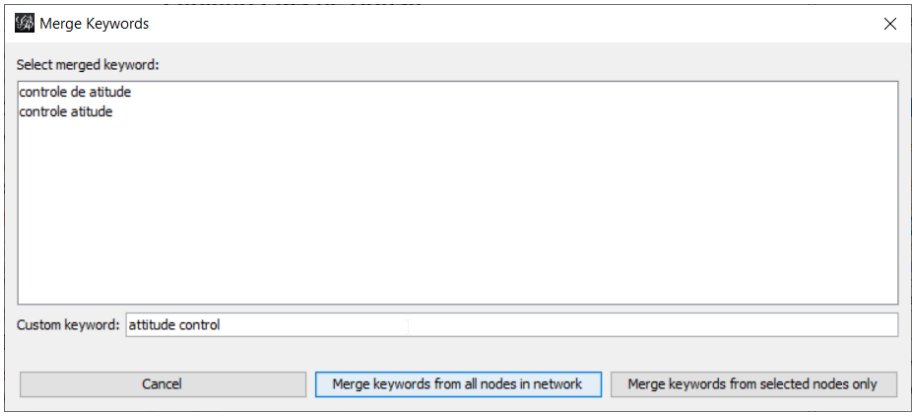


Figura 7.23 Especificação de uma palavra-chave nova

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:

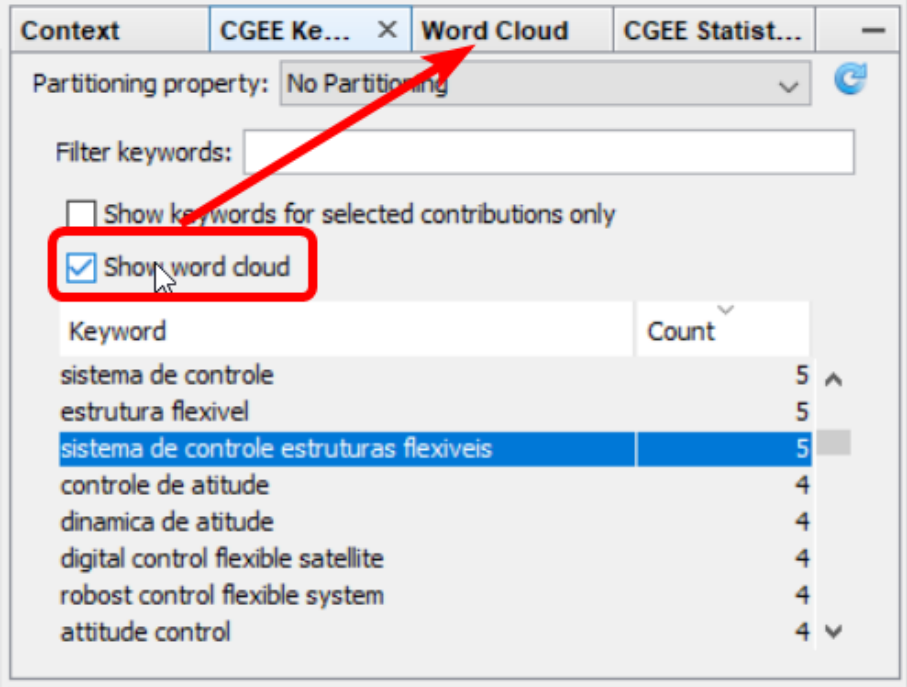


Figura 7.24 Habilitação da nuvem de palavras-chave

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

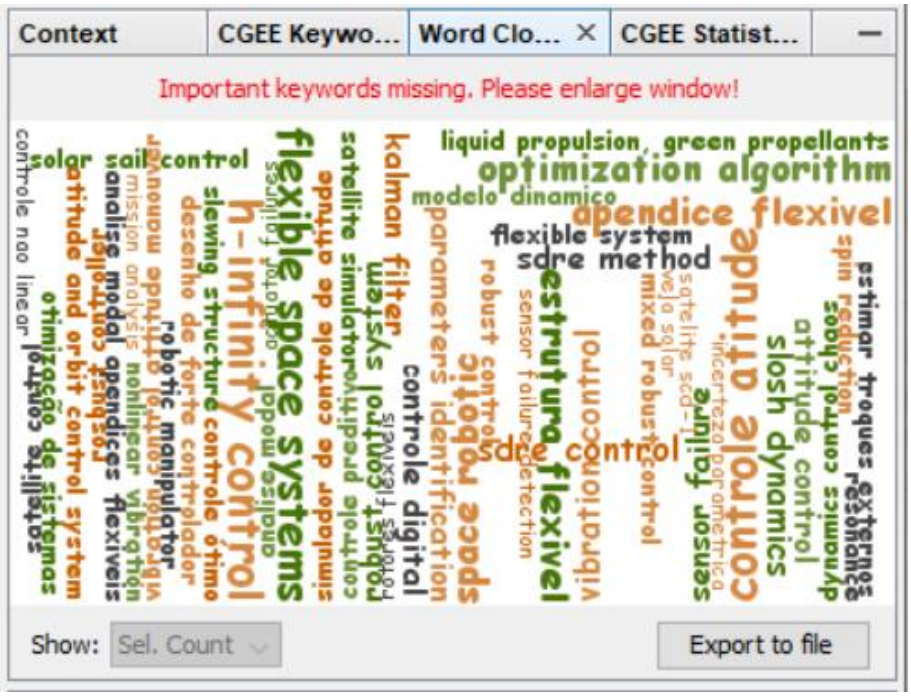


Figura 7.25 Nuvem de palavras-chave, com aviso de falta de precisão

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:

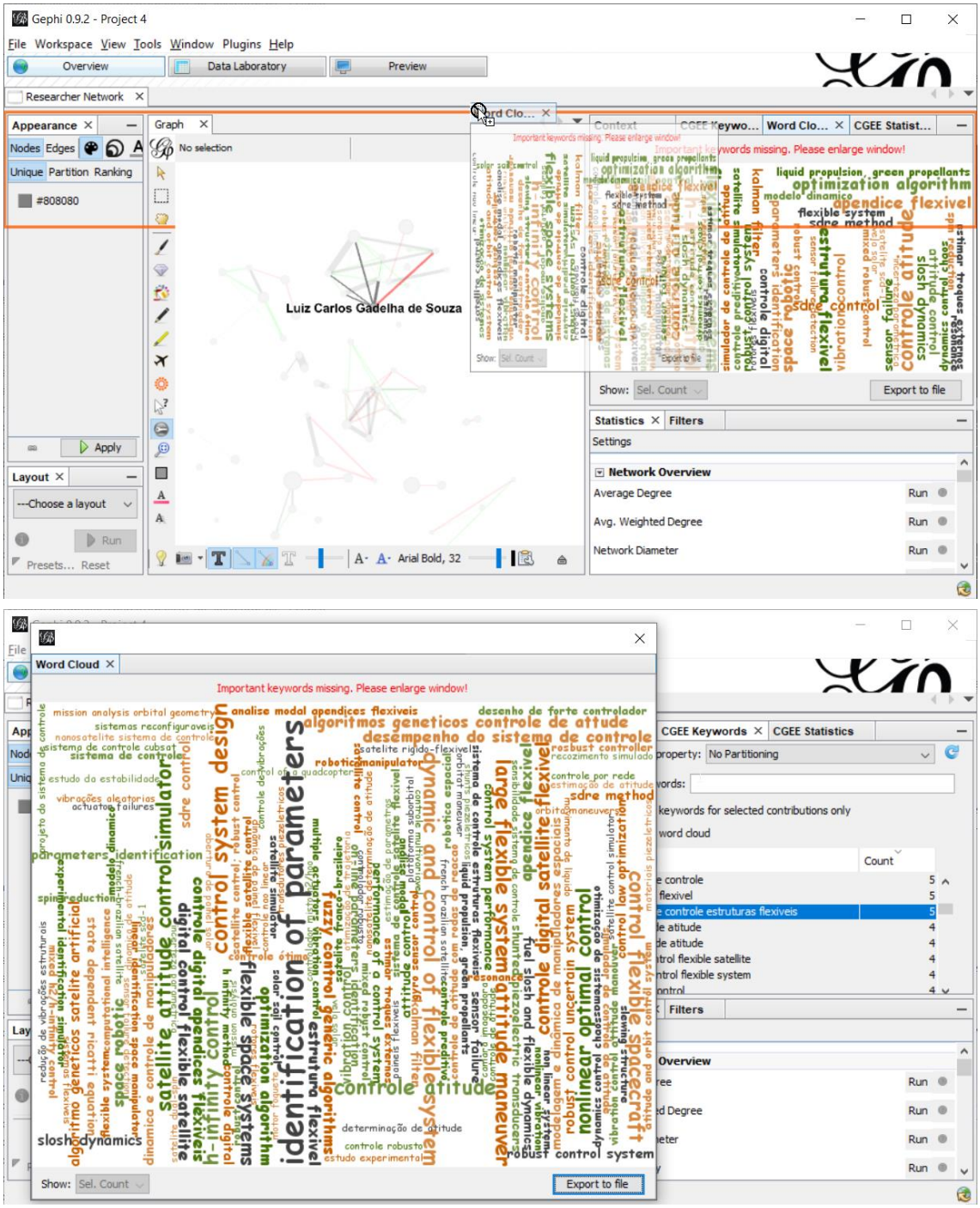


Figura 7.26 Separação da janela de nuvem de palavras

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 [7]. 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:

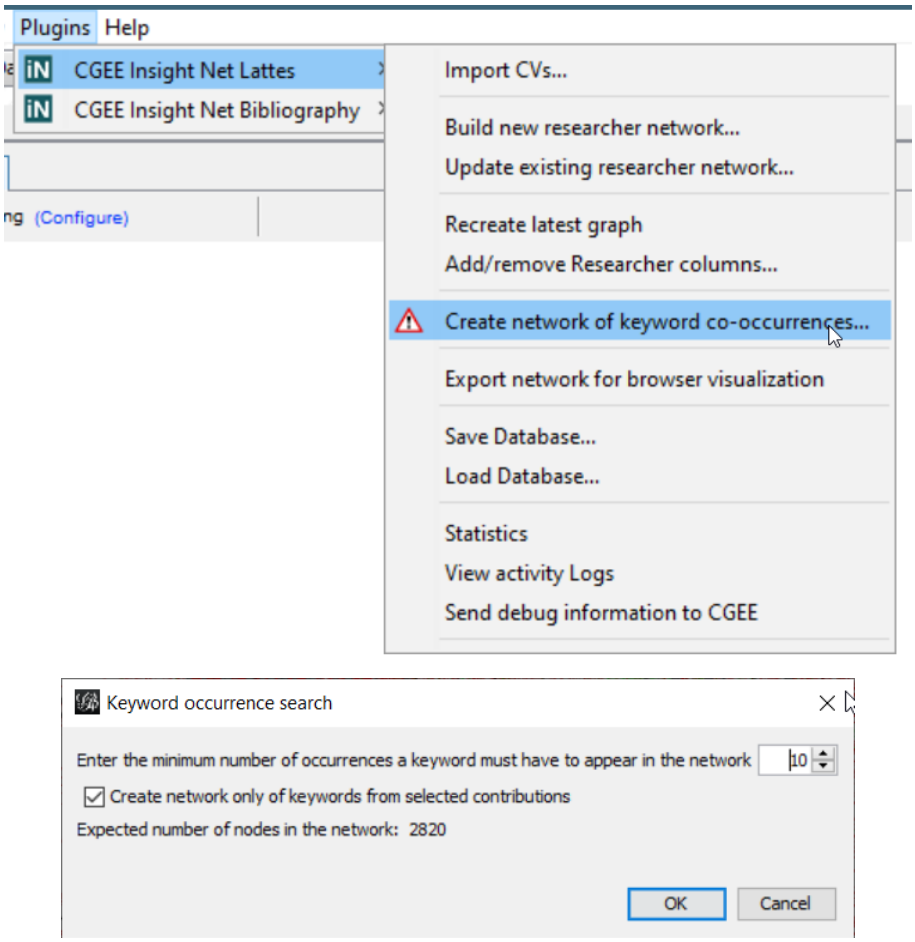


Figura 7.27 Funcionalidade para criar redes de palavras-chave

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).

7.6 Eliminação interativa de nós da rede e do banco de

[7] *Contribuições bibliográficas* são as referências bibliográficas importadas pelos módulos "BibTeX" e "Bibliografia genérica", bem como os artigos, capítulos de livros e trabalhos em eventos que constam nos Currículos Lattes.

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:

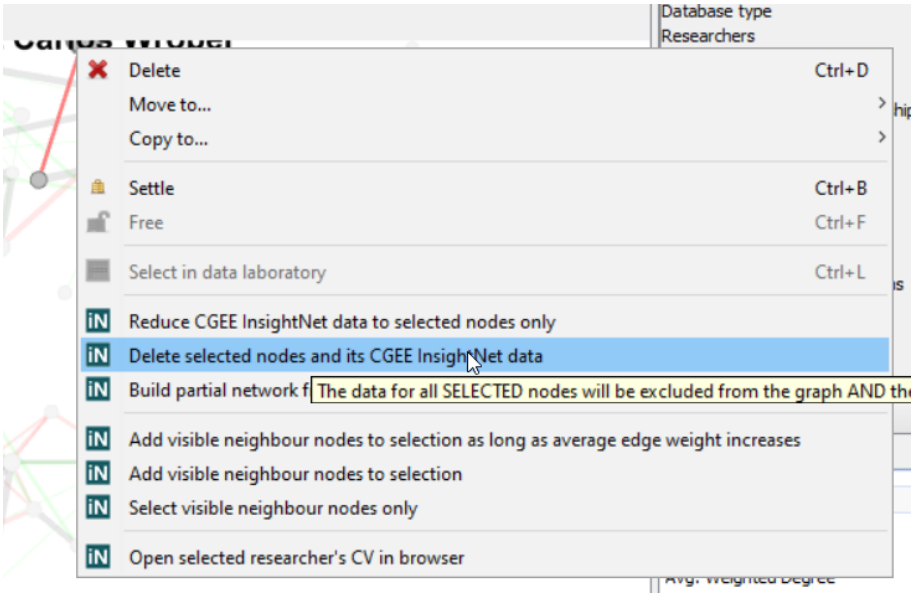


Figura 7.28 Eliminação interativa de nós do banco de dados na visualização do grafo

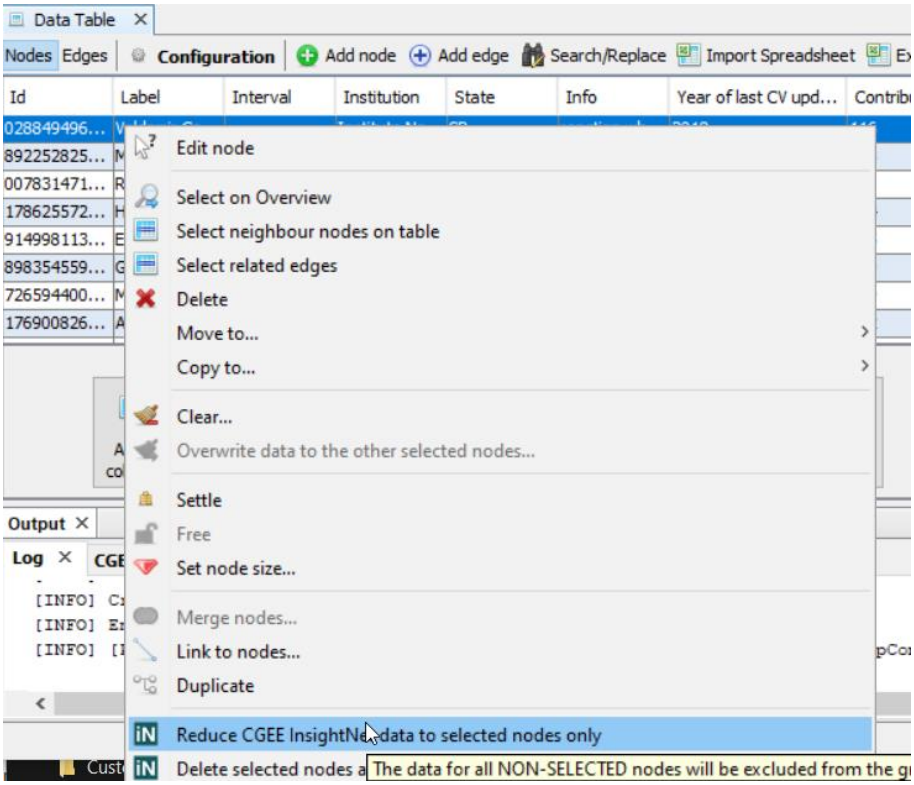


Figura 7.29 Eliminação interativa de nós do banco de dados no laboratório de dados

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 selecionados

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.

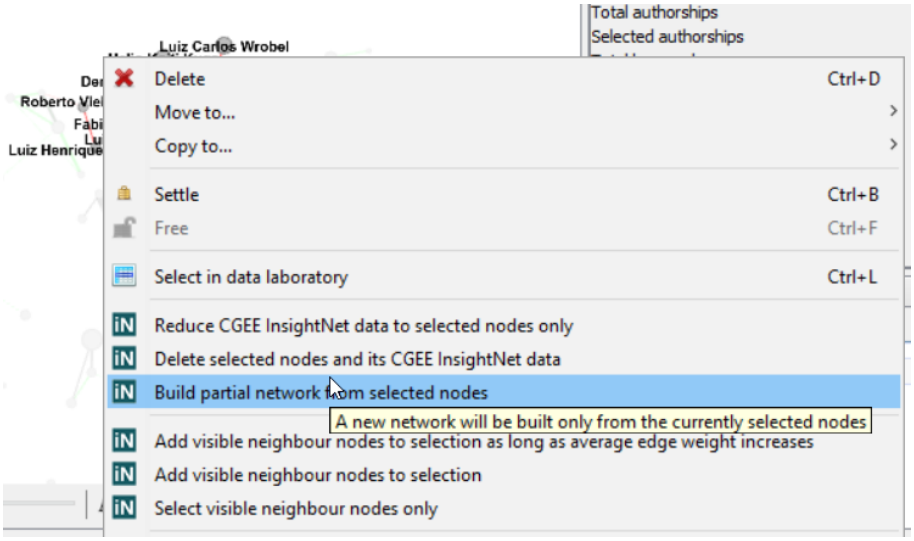


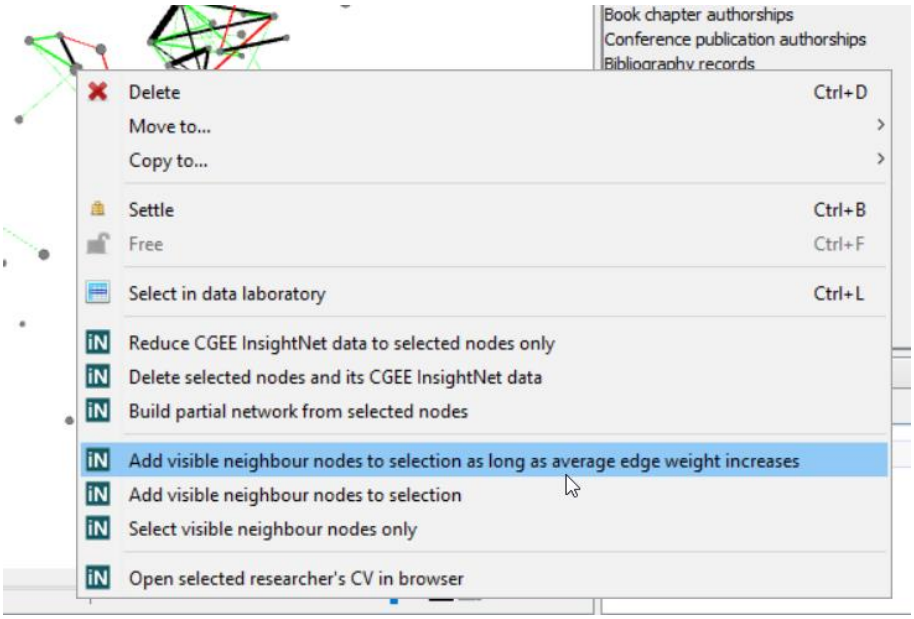
Figura 7.30 Criação de uma rede a partir do subconjunto de nós selecionados

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.

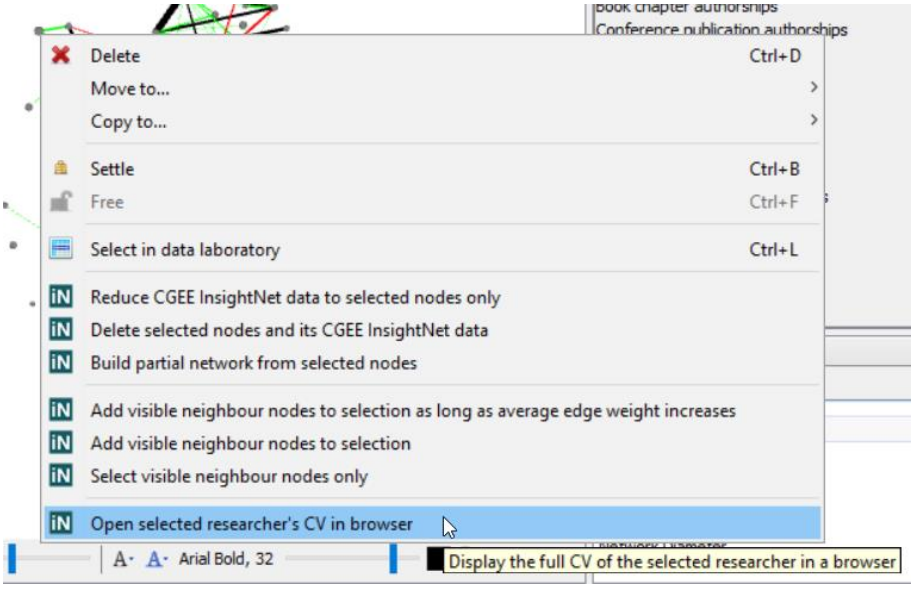


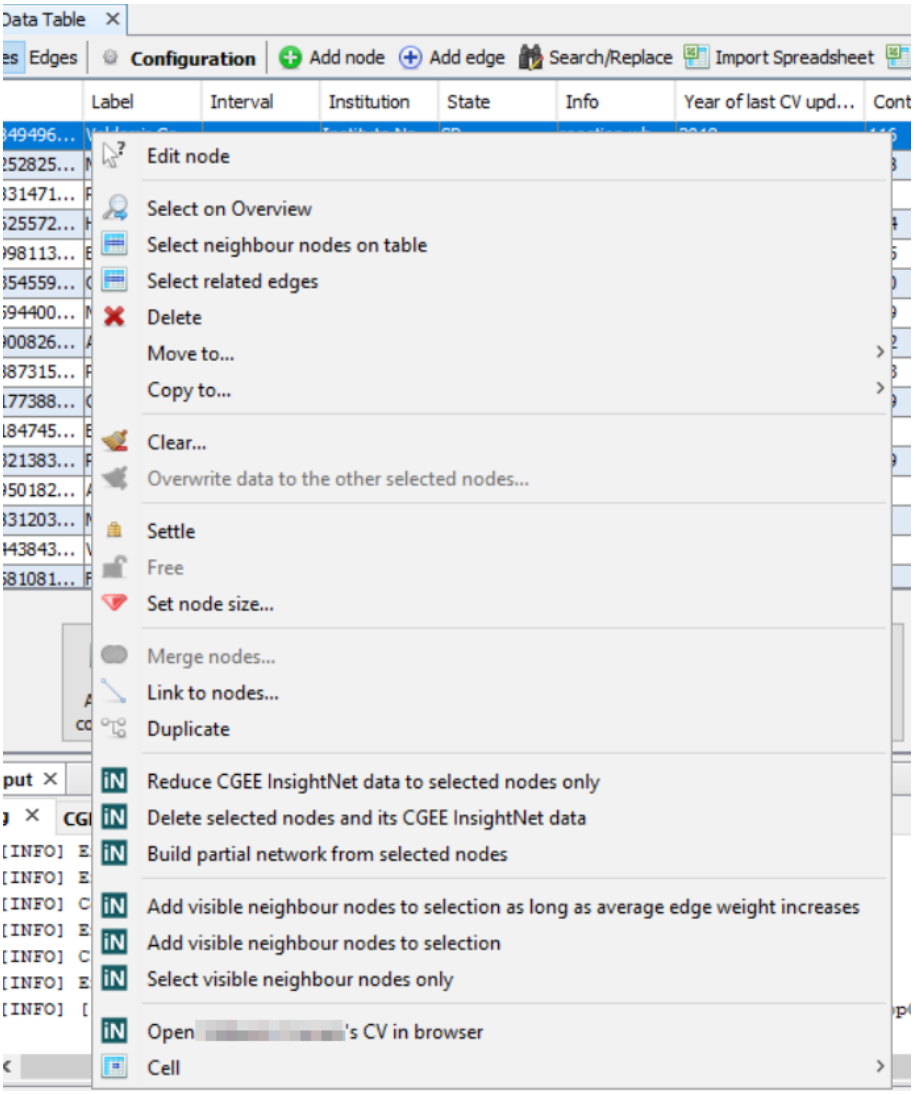
Já a funcionalidade "*Add visible neighbour nodes to selection as long as average edge weight increases*" repete o processo de adicionar vizinhos enquanto o peso média das arestas não diminua. Desta forma, o processo termina quando os nós adicionados não

agregam mais informações relevantes ao subconjunto de nós selecionados.

7.9 Visualização interativa do currículo de pesquisadores no browser

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...
Cont
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

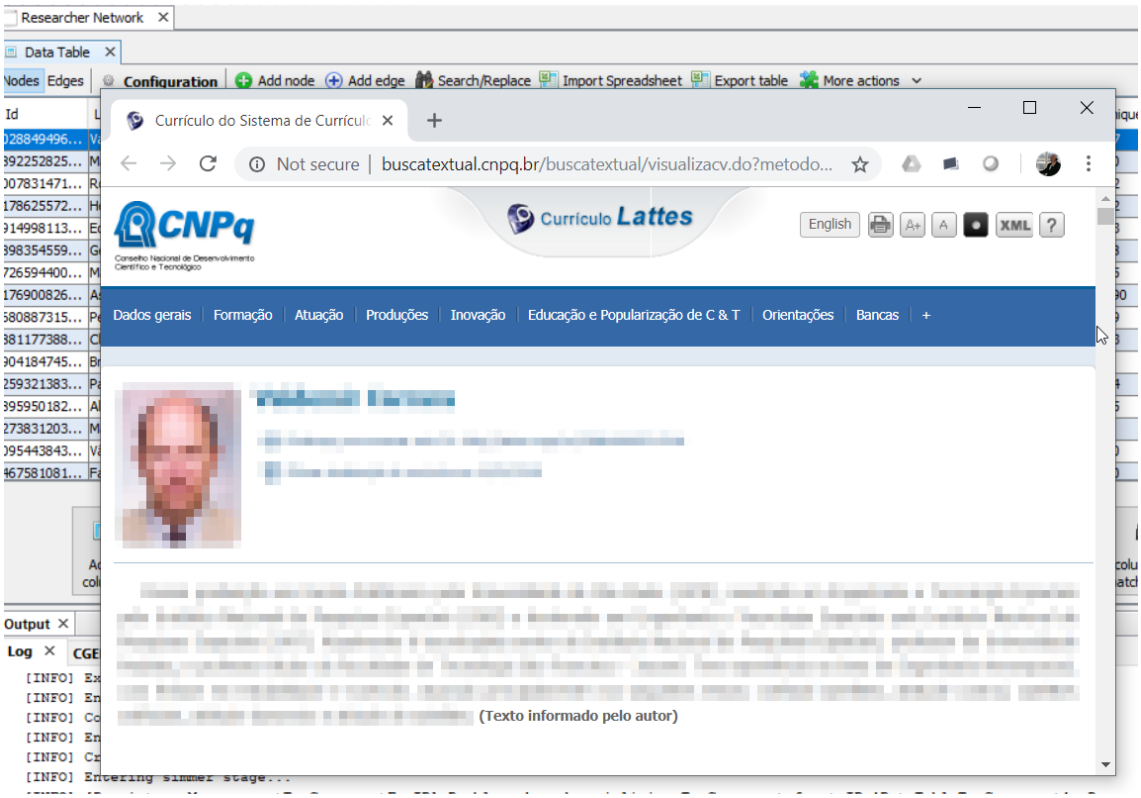
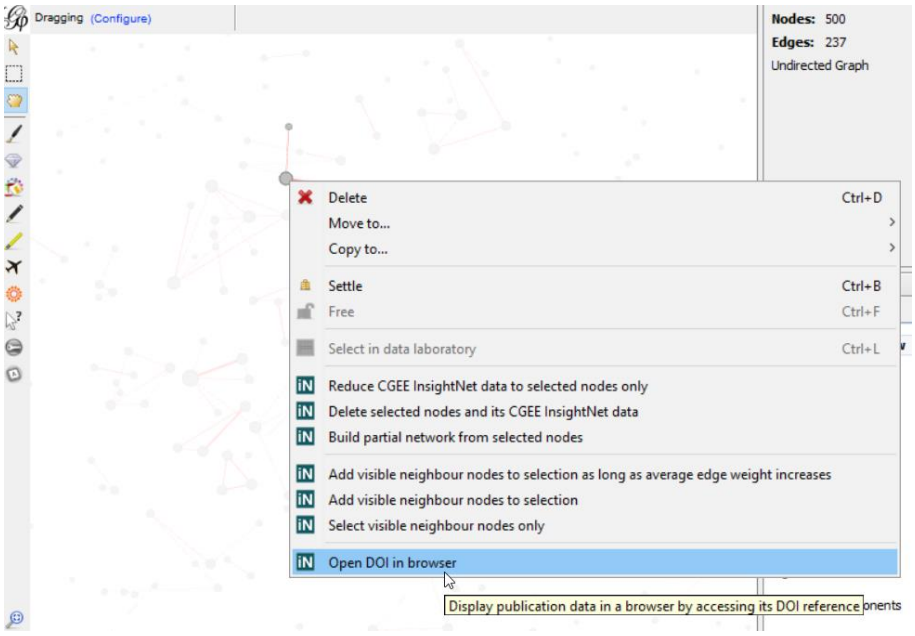
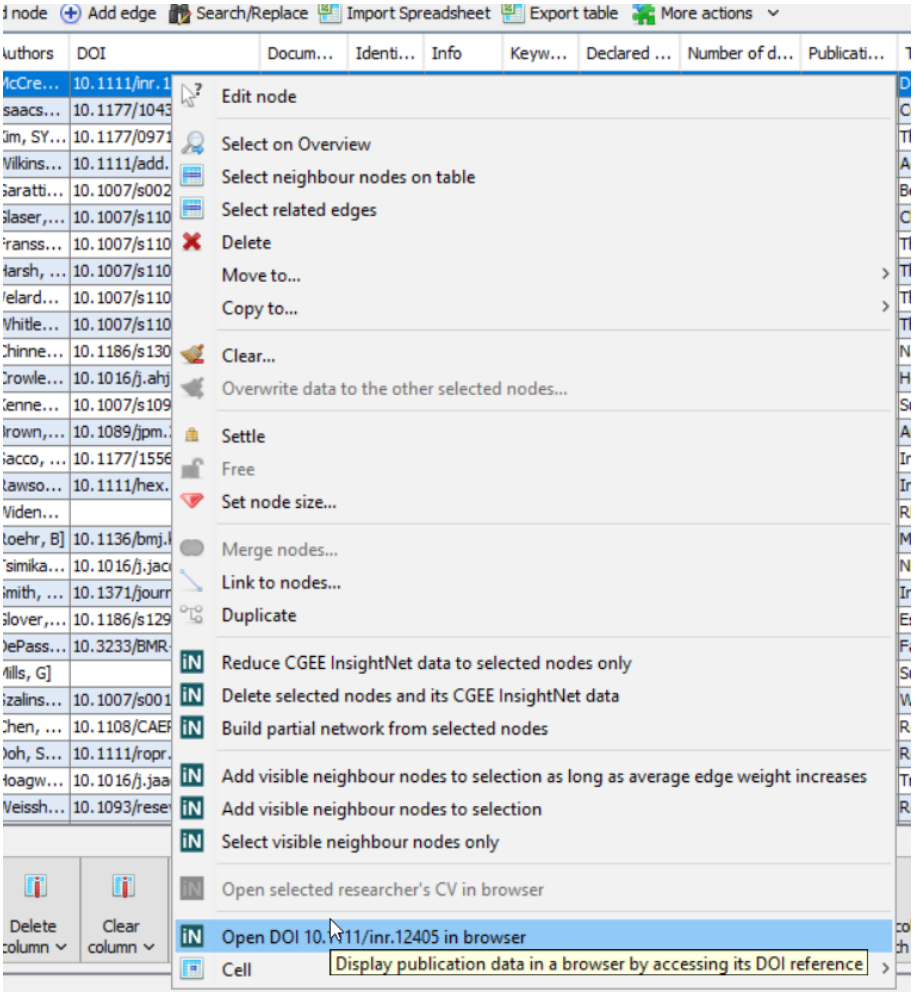


Figura 7.31 Visualização do currículo do pesquisador no Browser

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.




node
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
Cell
Display publication data in a browser by accessing its DOI reference
Delete column
Clear column

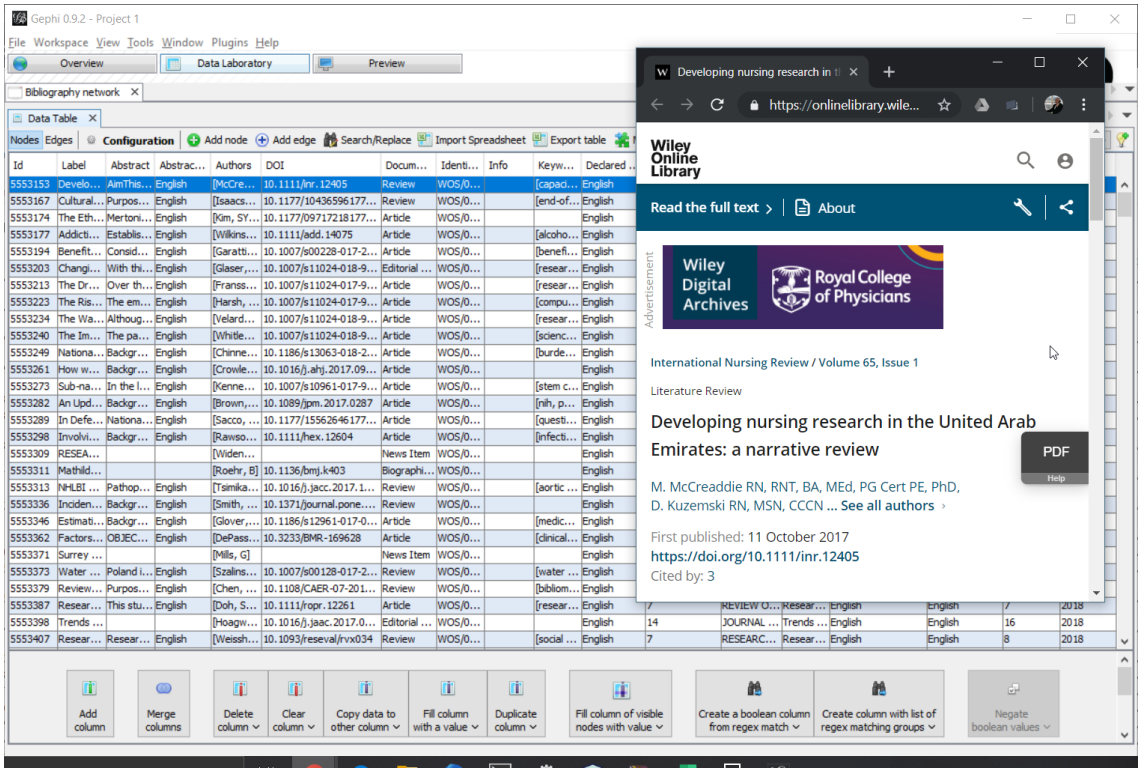


Figura 7.32 Visualização da referência bibliográfica por DOI no Browser

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:

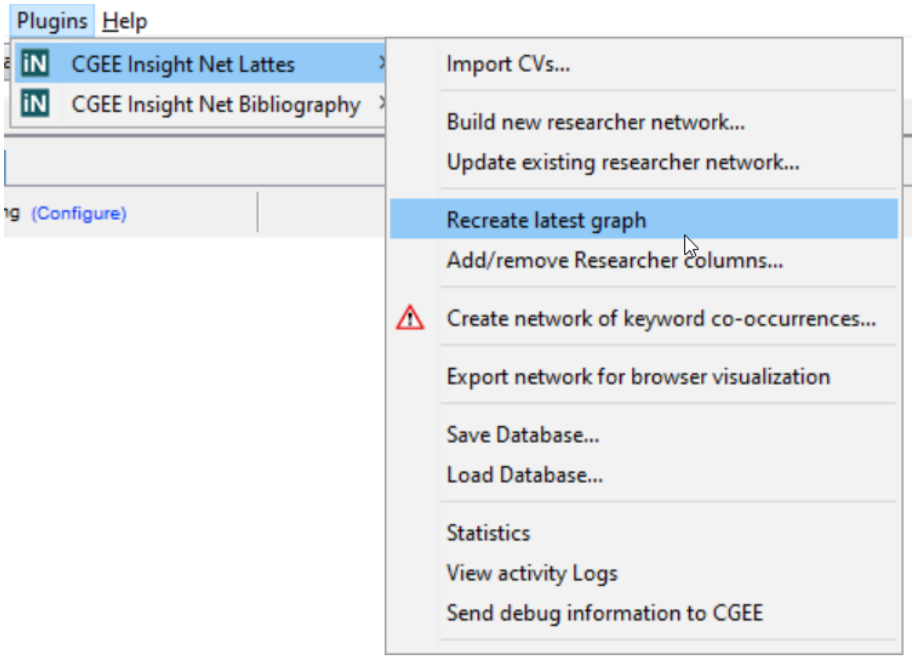


Figura 8.1 Recuperação do grafo

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:

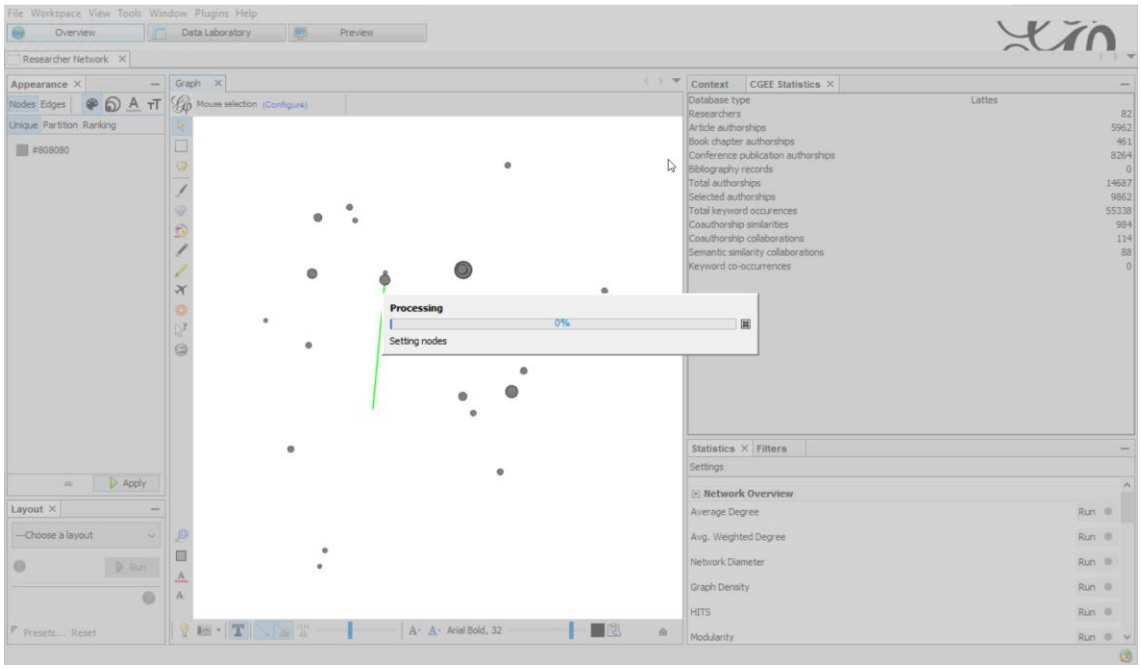


Figura 8.2 Recuperação do grafo

8.2 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* …:

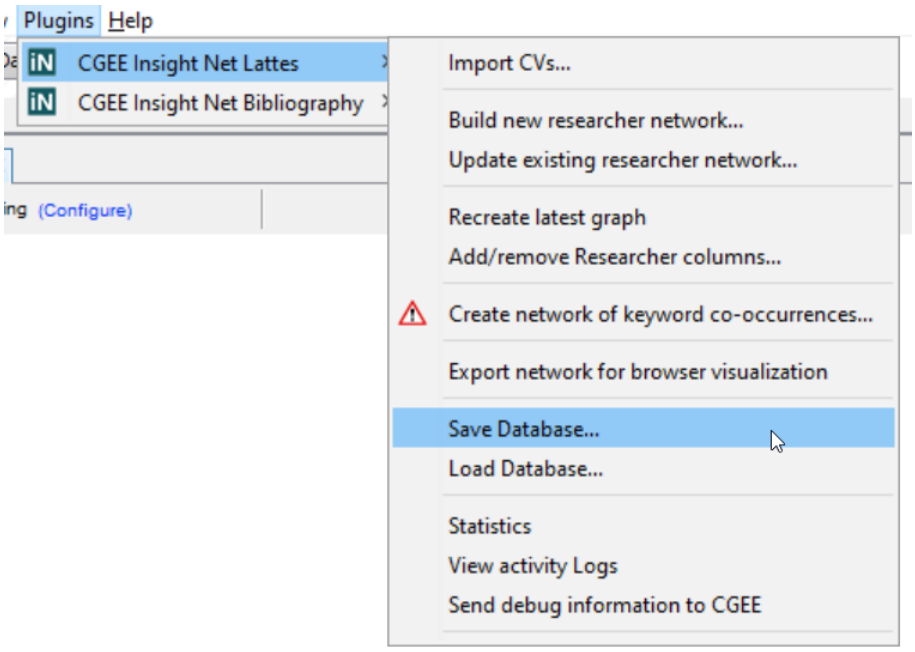


Figura 8.3 Funcionalidades de gravação e recuperação do banco de dados

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

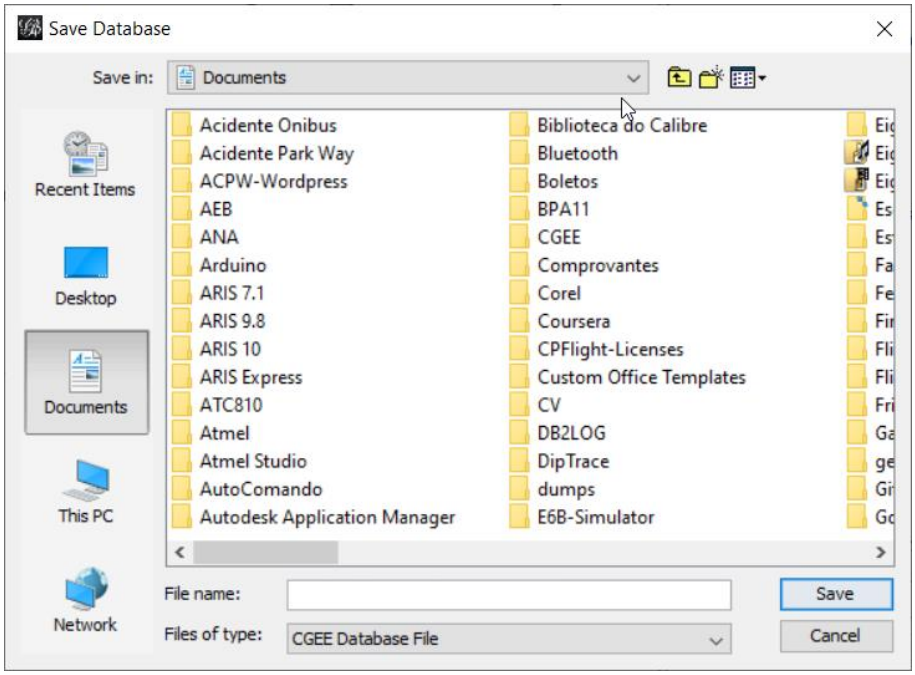


Figura 8.4 Especificação do arquivo de backup do banco de dados

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

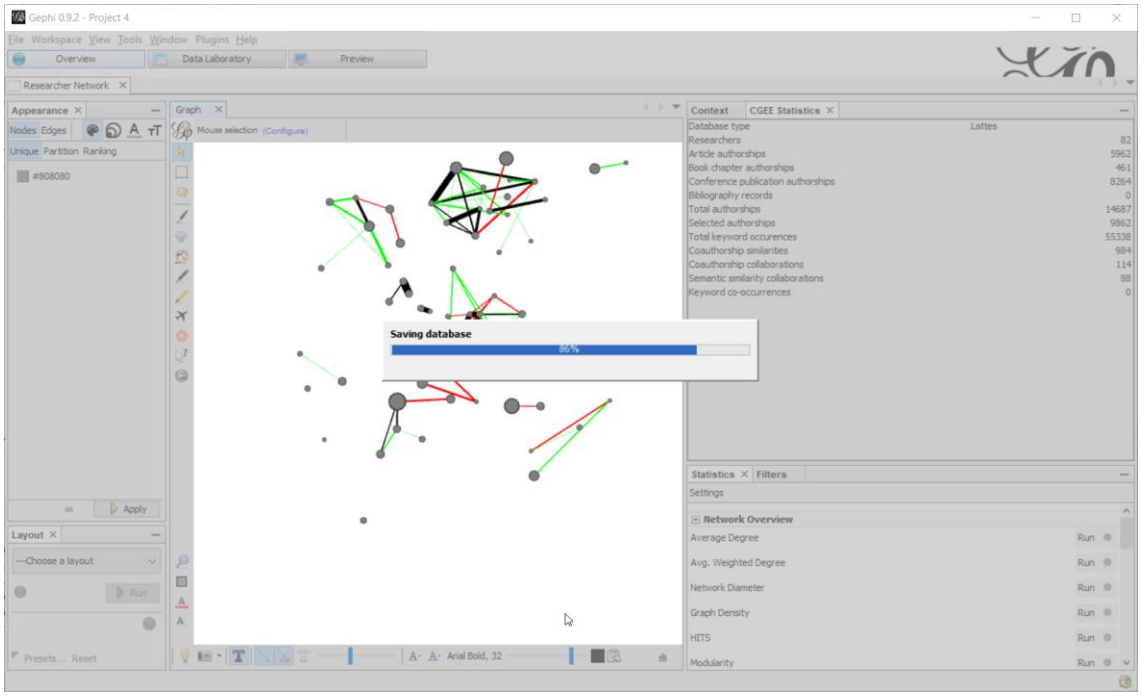
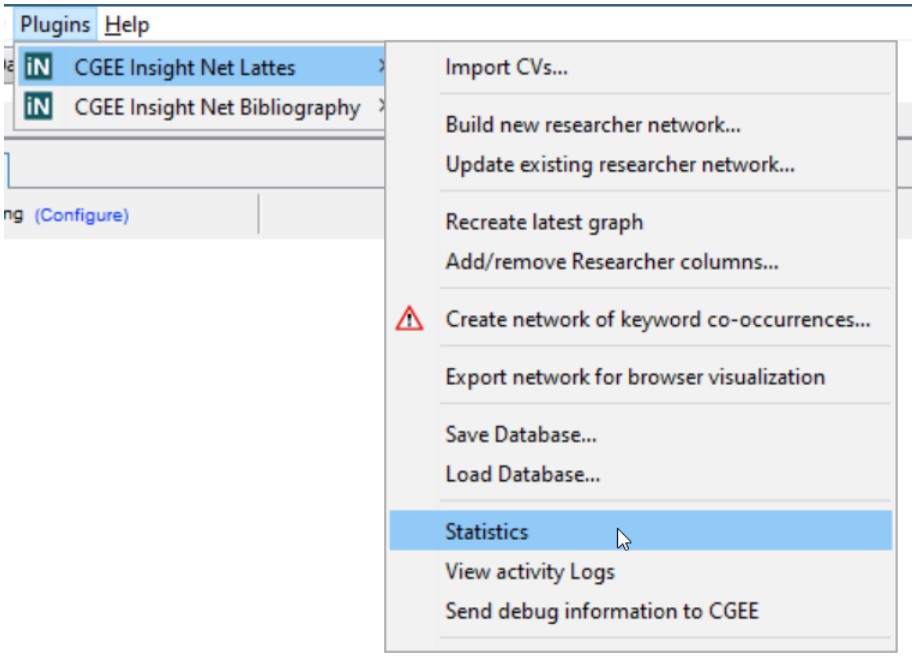


Figura 8.5 Geração do backup da base de dados

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:



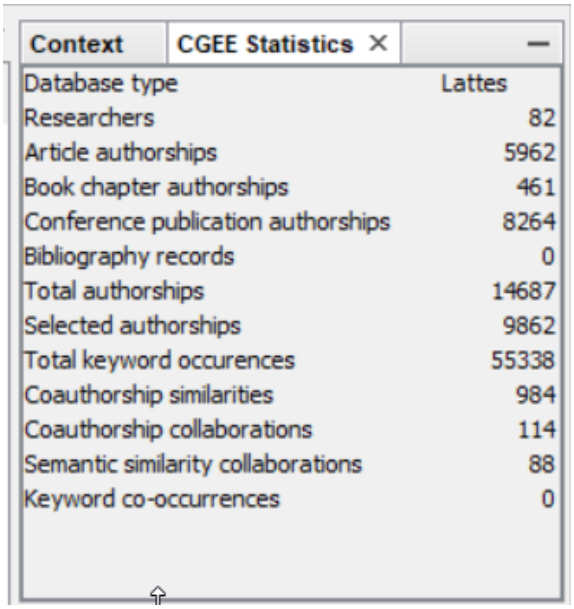
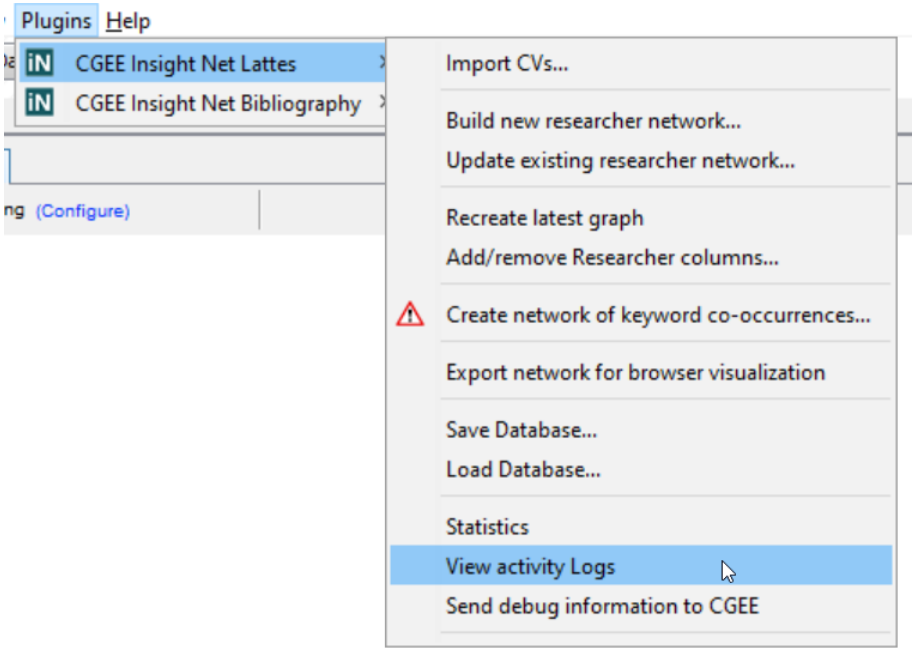


Figura 8.6 Janela de estatística

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:



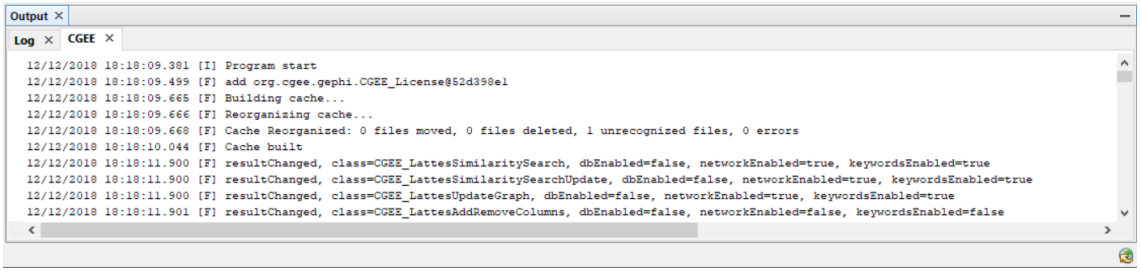


Figura 8.7 Protocolo de execução

A janela possui duas abas:

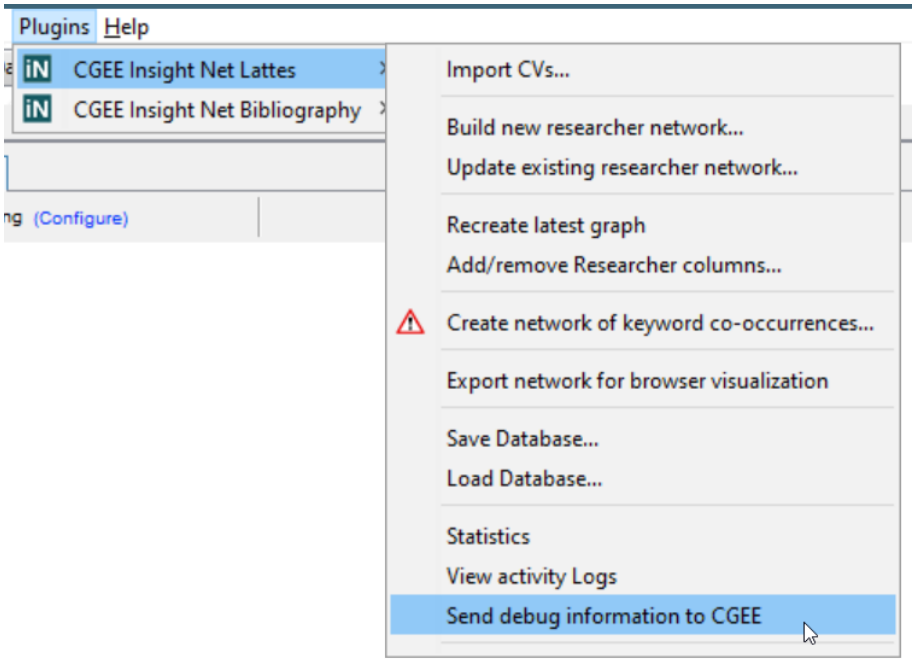
- Na aba "*Log*" são exibidas mensagens do proprio ambiente do Gephi, sem relação ao *CGEE Insight Net*
- Já a aba "CGEE" exibe os registros de protocolo de execução do *CGEE Insight Net*.

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*:



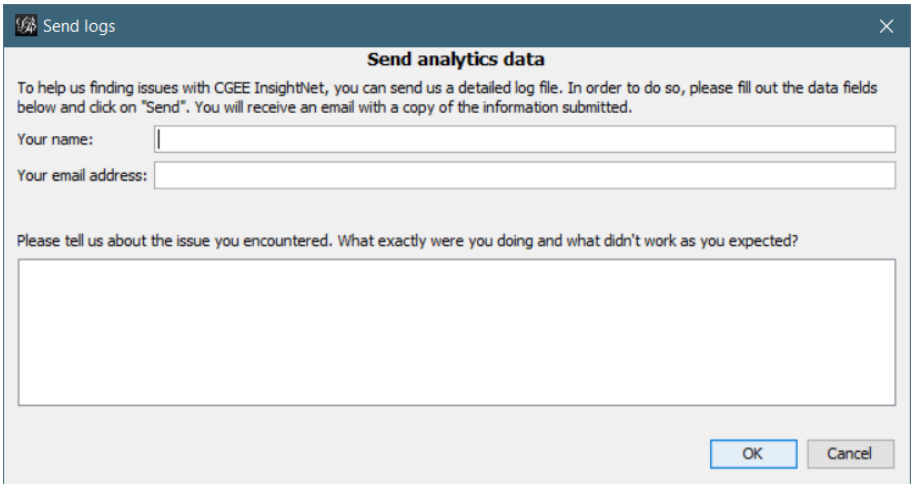


Figura 8.8 Enviar protocolos de execução

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:

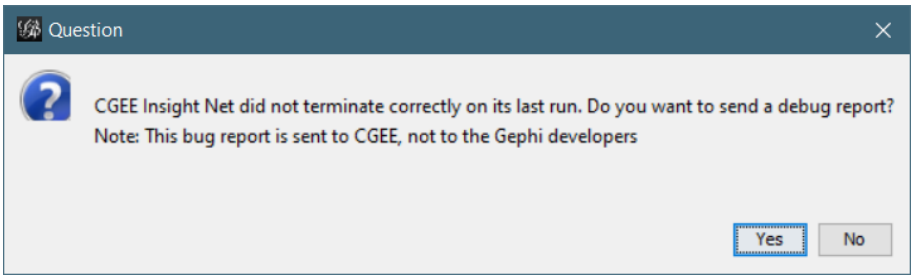


Figura 8.9 Diálogo que aparece depois da terminação forçada do Gephi

9 Informações adicionais

9.1 Especificação de formatos para o módulo de referências bibliográficas genéricas

O módulo de referências bibliográficas genéricas permite a importação de dados estruturados em diversos formatos a partir de arquivos descritivos desses formatos.

Os arquivos estão em formato JSON, em forma estruturada. Existem, três tipos de entidades nessa estrutura:

- *Descriptor* é a entidade raíz que descreve características globais dos arquivos a serem importadas. O *Descriptor* contém, entre outros, vários *Fields*.
- *Field* é a entidade que descreve um atributo do conjunto de dados a ser carregado, como título, ano de publicação ou palavras-chave. A definição de um *Field* pode incluir vários *Transformers*.
- Os dados de um *Field* podem ser tratados com um ou mais *Transformers*, que realizam operações genéricos.

Como exemplo, segue o formato de importação de dados genéricos de planilhas Excel. Os elementos específicos serão descritos nas seguintes seções:

```
{
    "name": "Generic",
    "description": "Generic [autodetect]",
```

```
    "lineFormat": "AUTO",
    "hasHeaderLine": "true",
    "addUnknownFieldsAsAttributes": "true",
    "createFields": "true",
    "fieldDefs": [
        {
            "name": "Identifier",
            "tag": "ID",
            "fieldType": "ID"
        },{
            "name": "DOI",
            "tag": "DOI",
            "transformers": [
                {
                    "transformerType": "LOWERCASE"
                }
            ],
            "fieldType": "DOI"
        },{
            "name": "Title",
            "tag": "Title",
            "fieldType": "TITLE"
        },{
            "name": "Abstract",
            "tag": "Abstract",
            "fieldType": "ABSTRACT"
        },{
            "name": "Authors",
            "tag": "Authors",
            "transformers": [
                {
                    "transformerType": "SPLIT",
                    "ifFormat": ["CSV", "TSV", "EXCEL"],
                    "matchRegex": ";¥¥s+"
                }
            ],
            "fieldType": "AUTHORS"
        },{
            "name": "Language",
            "tag": "Language",
            "fieldType": "LANGUAGE"
        },{
            "name": "Keywords",
            "tag": "Keywords",
            "transformers": [
                {
                    "transformerType": "SPLIT",
                    "ifFormat": ["CSV", "TSV", "EXCEL"],
                    "matchRegex": ";¥¥s+"
                }
            ],
            "fieldType": "KEYWORDS"
        },{
            "name": "Year",
            "fieldType": "YEAR",
```

```
            "tag": "Year",
            "transformers": [
                {
                    "transformerType": "NUMBER"
                }
            ]
        },{
            "name": "Document type",
            "tag": "Document Type",
            "fieldType": "TYPE"
        },{
            "name": "Source Title",
            "tag": "Source title",
            "fieldType": "SOURCE"
        }
    ]
}
```

9.1.1 Descriptor

A entidade *descriptor* é a entidade raiz do arquivo de descrição de formato de dados e possui os seguintes campos:

Tabela 9.1 Campos do Descriptor

Campo	Significado
`name`	Nome do formato, utilizado para referências internas
`description`	Breve descrição do formato, utilização interna
`lineFormat`	Formato das linhas. Aceita os valores `AUTO` (identificação automática), `CSV` (separação por vírgula ou outro caractere definido pelo campo `separationChar`), `TSV` (separação por caractere TAB `0x09`), `TAGGED` (formato RIS) ou `EXCEL` (planilha Excel®).
`separationChar`	Caso especificado, é o caractere que separa os campos no formato `CSV`. Se não for especificado, é utilizada a vírgula.
`lineFormatType`	Determina o tipo de arquivo para o formato `TAGGED`. Este formato possui variações para arquivos do "*Web of Science*" e do "*Scopus*". Assim, o campo aceita os valores `Scopus` e `Web of Science`.
`hasHeaderLine`	Campo booleano (valores `true` ou `false`), que especifica se a primeira linha do arquivo contém os nomes das colunas. Relevante para arquivos nos formatos `TSV`, `CSV` e `EXCEL`.
`columnList`	Para formatos `TSV`, `CSV` e `EXCEL`, este campo permite especificar uma lista com nomes das colunas, caso o campo `hasHeaderLine` tiver o valor `false`.
`addUnknownFieldsAsAttributes`	Campo booleano que define que qualquer coluna de dados que possui um nome, mas que não foi definida explicitamente no *descriptor* será importada como atributo com o nome da coluna.
`include`	Lista de arquivos de *descriptors* que serão incluidos no

	processamento
`fieldDefs`	Lista de *Fields*, que descrevem os atributos da importação. Ver Seção 9.1.2
`ignoreRisTags`	Lista com nomes dos *Fields* que serão ignorados na importação de dados no formato `TAGGED`

9.1.2 Field

A entidade *Field* descreve o conteúdo e o tratamento dos campos individuais de dados durante e depois da importação. A lista dde *Field*s consta no atributo ``fieldDefs`` da entidade *descriptor*.

Tabela 9.2 Campos do Field

Campo	Significado
`name`	Nome do campo. Este nome será o nome do atributo do nó no Gephi, ou seja, o nome da coluna no laboratório de dados.
`fieldType`	Tipo do campo de dados, definindo o tratamento no Gephi. Os seguintes tipos de campos são definidos: • `IGNORE`: O campo não será importado no Gephi • `ID`: O campo será importado como identificador do nó no Gephi • `DOI`: Document Object Identifier, usado para desduplicar documentos • `TITLE`: Título da publicação, usado na busca de similaridade • `ABSTRACT`: Resumo da publicação, usado na busca de similaridade • `AUTHORS`: Lista de autores • `KEYWORDS`: Lista de palavras-chave da publicação • `KEYWORDS_ADDITIONAL`: Lista de palavras-chave adicionais da publicação • `YEAR`: Ano de publicação, usado no filtro de busca de similaridade • `LANGUAGE`: Idioma da publicação. Valores permitidos: “english”, “portuguese”, “inglês” e “português” • `SOURCE`: Nome da fonte da publicação (revista, livro etc.) • `TYPE`: Tipo de publicação (texto livre) • `ATTRIBUTE`: O campo será importado como atributo genérico, sem tratamento específico
`tag`	Este campo contém o nome do *Field* como ele aparece no cabeçalho dos arquivosCSV e TSV e é usado para identificar o *field* na lista de dados durante a importação de um desses formatos.
`risTag`	Este campo contém o nome do *Field* como ele aparece nos arquivos `RIS` (formato `TAGGED`). O valor é usado para identificar o *Field* durante a importação neste formato.
`multipleItems`	Campo booleano (valores `true` e `false`) que especifica que o conteúdo do *Field* é uma lista de valores.
`transformers`	Lista de *transformers*, que definem um tratamento do(s) valor(es) durante a importação dos dados. Ver Seção 9.1.3
`insertTransformers`	Este campo pode ter os valores `BEFORE` ou `AFTER` e define se na

	os *Transformers* deste campo serão executados antes ou depois de *Transformers* já existentes.

9.1.3 Transformer

A entidade *Transformer* define tratamentos dos dados durante a importação. Cada *Field* pode ter uma lista de *Transformers* em seu campo `transformers`.

Tabela 9.3 Campos do Transformer

Campo	Significado
`transformerType`	Tipo do Transformer. Os seguintes tipos são definidos e descritos em seguida: ● `ALWAYS`: Aplicar expressão regular no valor do campo ● `CONDITIONAL`: Aplicar expressão regular se o valor do campo confere com outra expressão regular ● `SPLIT`: Dividir o valor de campo em uma lista, usando expressão regular como critério de divisão ● `SELECT`: Selecionar um item específico da lista de valores ● `FILTER`: Eliminar valores que não conferem com uma expressão regular ● `DEDUP`: Eliminar valores duplicados da lista de valores ● `NUMBER`: Transforma valor do campo em um valor numérico ● `LOWERCASE`: Transformar texto do valor em letras minúsculas ● `FORMATINFO`: *Tag* adicional usado em arquivos do formato `TAGGED` do Scopus®
`ifFormat`	O *Transformer* é executado apenas se o formato do arquivo consta na lista de formatos especificado neste campo. Exemplo: `ifFormat: ["CSV","TSV", "EXCEL"]`

`conditionRegex`
`matchRegex`
`substitutionRegex`
	Expressões regulares usados no transformador. O conteúdo depende do tipo de tipo de transformador e é descrito nas seções seguintes.

9.1.3.1 *Transformer* `ALWAYS`

Este *transformer* substitui todas os ocorrências da expressão regular `matchRegex` no(s) valor(es) do campo pela expressão regular `substitutionRegex`. Grupos da primeira expressão regular são referenciados na segunda com o símbolo `$`. O seguinte exemplo demonstra parte do tratamento do país nos dados do Scopus:

```
[...]
            },{
                "transformerType": "ALWAYS",
                "matchRegex": "^(.*)([.])$",
                "substitutionRegex": "$1"
            },{
                "transformerType": "ALWAYS",
                "matchRegex": "(.*)([Cc][Hh][Ii][Nn][Aa])(.*)",
                "substitutionRegex": "$2"
            },{
                "transformerType": "ALWAYS",
                "matchRegex": "(.*)([Cc][Aa][Nn][Aa][Dd][Aa])(.*)",
                "substitutionRegex": "$2"
            },{
```

```
                "transformerType": "ALWAYS",
                "matchRegex": "(.*?)(¥¥s+)(.+)(¥¥s+)(.+)(¥¥s+)(.*)",
                "substitutionRegex": "$7"
            },{
                "transformerType": "ALWAYS",
                "matchRegex":
"^[A-Za-z][A-Za-z]¥¥s[0-9][0-9][0-9][0-9][0-9][-][0-9][0-9][0-9][0-9]$",
                "substitutionRegex": "United States"
            },{
                "transformerType": "ALWAYS",
                "matchRegex":
"^[A-Za-z][A-Za-z]¥¥s[0-9][0-9][0-9][0-9][0-9]$",
                "substitutionRegex": "United States"
            },{
                "transformerType": "ALWAYS",
                "matchRegex": "^[^0-9]*[0-9][0-9][0-9][0-9][0-9]$",
                "substitutionRegex": "United States"
            },{
                "transformerType": "ALWAYS",
                "matchRegex": "^[A-Za-z][A-Za-z]$",
                "substitutionRegex": "United States"
            },{
[...]
```

9.1.3.2 *Transformer* `CONDITIONAL`

Este *transformer* funciona análogo ao *transformer* `ALWAYS`, mas permite condicionar a substituição. Ele é executado apenas se o valor de entrada confere com a expressão regular `conditionRegex`. Se este campo não foi definido no *transformer*, ele se comporta igual ao *transformer* `ALWAYS`.

9.1.3.3 *Transformer* `SPLIT`

Este *transformer* quebra um único valor em uma lista (um vetor) de valores. Como critério de divisão, a expressão regular `matchRegex` é utilizado. Segue, como exemplo, o tratamento de autores nos dados do Scopus®, apenas para dados nos formatos `CSV`,``TSV`` e `EXCEL`:

```
[...]
    },{
        "name": "Authors",
        "fieldType": "AUTHORS",
        "tag": "Authors",
        "risTag": "AU",
        "transformers": [
            {
                "transformerType": "SPLIT",
                "ifFormat": ["CSV", "TSV", "EXCEL"],
                "matchRegex": ",¥¥s+"
            }
        ]
    },{
[...]
```

9.1.3.4 *Transformer* `SELECT`

Este *transformer* reduz uma lista (um vetor) de valores para um único elemento (escalar), definido pela posição na lista. A posição do elemento selecionado é definido como valor

numérico no campo `matchRegex`. O valor `0` seleciona o primeiro elemento da lista, o valor `1` o segundo etc. Valores negativos selecionam elementos a partir do fim da lista: `-1` é o último elemento da lista, `-2` o penúltimo etc:

```
[...]
},{
    "name": "Organizations",
    "fieldType": "ATTRIBUTE",
    "tag": "C1",
    "multipleItems": "true",
    "transformers": [
        {
            "transformerType": "ALWAYS",
            "matchRegex": "¥¥[.+?¥¥]¥¥s+",
            "substitutionRegex": ""
        },{
            "transformerType": "SPLIT",
            "ifFormat": ["CSV", "TSV", "EXCEL"],
            "matchRegex": ";¥¥s+"
        },{
            "transformerType": "SPLIT",
            "matchRegex": ",¥¥s*"
        },{
            "transformerType": "SELECT",
            "matchRegex": "0"
        },{
            "transformerType": "DEDUP"
        }
    ]
},{
[...]
```

9.1.3.5 *Transformer* `FILTER`

O *transformer* `FILTER` elimina todos os valores de uma lista de valores que não conferem com a expressão regular `matchPattern`:

```
[...]
},{
    "name": "ISSN",
    "fieldType": "ATTRIBUTE",
    "tag": "ISSN",
    "risTag": "SN",
    "transformers": [
        {
            "transformerType": "SPLIT",
            "matchRegex": ";¥¥s+"
        },{
            "transformerType": "FILTER",
            "ifFormat": [ "TAGGED" ],
            "matchRegex": "(.*)¥¥s+¥¥(ISSN¥¥)"
        },{
            "transformerType": "CONDITIONAL",
            "matchRegex": "(.*)¥¥s+¥¥(ISSN¥¥)",
            "substitutionRegex": "$1"
        }
```

```
    ]
},{
[...]
```

9.1.3.6 *Transformer* DEDUP

O *transformer* DEDUP elimina os valores duplicados de uma lista de valores:

```
[...]
},{
    "name": "Organizations",
    "fieldType": "ATTRIBUTE",
    "tag": "C1",
    "multipleItems": "true",
    "transformers": [
        {
            "transformerType": "ALWAYS",
            "matchRegex": "¥¥[.+?¥¥]¥¥s+",
            "substitutionRegex": ""
        },{
            "transformerType": "SPLIT",
            "ifFormat": ["CSV", "TSV", "EXCEL"],
            "matchRegex": ";¥¥s+"
        },{
            "transformerType": "SPLIT",
            "matchRegex": ",¥¥s*"
        },{
            "transformerType": "SELECT",
            "matchRegex": "0"
        },{
            "transformerType": "DEDUP"
        }
    ]
},{
[...]
```

9.1.3.7 *Transformer* NUMBER

O *transformer* NUMBER interpreta o conteúdo do *field* como valor numérico:

```
[...]
},{
    "name": "Year",
    "fieldType": "YEAR",
    "tag": "PY",
    "transformers": [
        {
            "transformerType": "NUMBER"
        }
    ]
},{
[...]
```

9.1.3.8 *Transformer* LOWERCASE

O *transformer* LOWERCASE substitui todas as letras maiúsculas no valor do *field* por letras minúsculas:

```
[...]
},{
    "name": "DOI",
    "fieldType": "DOI",
    "tag": "DOI",
    "risTag": "DO",
    "transformers": [
        {
            "transformerType": "LOWERCASE"
        }
    ]
},{
[...]
```

9.1.3.9 *Transformer* FORMATINFO

Este *transformer* é utilizado apenas em dados do formato TAGGED com lineFormatType SCOPUS. Neste formato, existem campos de dados que utilizam o mesmo *tag* N1. O valor do campo inicia com um texto específico que determina de que campo efetivamente se trata:

```
N1  - Conference code: 121263
N1  - Export Date: 15 March 2017
N1  - Funding details: MOE, Ministry of Education
N1  - Funding details: NRF-2013R1A1A2011601, MOE, Ministry of Education
N1  - Funding details: NRF, National Research Foundation of Korea
N1  - Funding text: This research was supported by Basic Science Research Program
through
the National Research Foundation of Korea (NRF) funded by the Ministry of Education
(NRF-2013R1A1A2011601)
and the Human Resource Training Program for Regional Innovation and Creativity
through the Ministry of
Education and National Research Foundation of Korea (NRF-2015H1C1A1035548).
```

Para poder distinguir esses casos, o *transformer* FORMATINFO permite especificar o texto inicial que determina o tipo de campo:

```
[...]
},{
    "name": "CODEN",
    "fieldType": "ATTRIBUTE",
    "tag": "CODEN",
    "risTag": "N1",
    "transformers": [
        {
            "transformerType": "FORMATINFO",
            "matchRegex": "CODEN:"
        }
    ]
},{
    "name": "Export date",
    "fieldType": "ATTRIBUTE",
    "risTag": "N1",
    "transformers": [
        {
            "transformerType": "FORMATINFO",
            "matchRegex": "Export Date:"
```

```
        }
    ]
},{
    "name": "Funding Details",
    "fieldType": "ATTRIBUTE",
    "tag": "Funding Details",
    "risTag": "N1",
    "transformers": [
        {
            "transformerType": "FORMATINFO",
            "matchRegex": "Funding details:"
        },{
            "transformerType": "SPLIT",
            "ifFormat": ["CSV", "TSV", "EXCEL"],
            "matchRegex": ";¥¥s+"
        }
    ]
},{
    "name": "Funding Text",
    "fieldType": "ATTRIBUTE",
    "tag": "Funding Text",
    "risTag": "N1",
    "transformers": [
        {
            "transformerType": "FORMATINFO",
            "matchRegex": "Funding text:"
        }
    ]
},{
    "name": "Tradenames",
    "fieldType": "ATTRIBUTE",
    "tag": "Tradenames",
    "risTag": "N1",
    "transformers": [
        {
            "transformerType": "FORMATINFO",
            "matchRegex": "Tradenames:"
        }
    ]
},{
[...]
```

9.2 Compilação do *CGEE Insight Net*

O *CGEE Insight Net* é composto de dois componentes básicos:

- A biblioteca `InsightNetLibrary`, que permite executar as operações a partir da linha de comando (veja Seção 9.3)
- O *plugin* `CGEE Insight Net 3.0`, que disponibiliza a interface de usuário para realizar as operações da linha de comando no *Gephi* e acrescenta componentes de filtros, estatísticas e análise de palavras-chave.

Os dois componentes são realizadas como módulos *Java* e utilizam a ferramenta *Maven* para facilitar o processo de compilação.

O código fonte para ambos os módulos consta no repositório *SVN* do CGEE, junto com arquivos de configuração da plataforma de desenvolvimento *Netbeans*, que permitem abrir

os projetos diretamente neste ambiente de desenvolvimento.

9.2.1 Biblioteca `InsightNetLibrary`

A biblioteca `InsightNetLibrary` possui um arquivo `pom.xml` que descreve todo o processo de compilação. O comando `mvn install` não requer configurações adicionais e gera o arquivo `InsightNetLibrary-<versão>.jar`.

Deve ser observado que o número de versão da biblioteca `InsightNetLibrary` deve ser o mesmo do *plugin* `CGEE Insight Net 3.0`.

9.2.2 *Plugin* `CGEE Insight Net 3.0`

A compilação do *Plugin* `CGEE Insight Net 3.0` ocorre de forma indireta, já que este *plugin* faz parte da infra-estrutura do *Gephi 0.9*. Conforme descrito no site de desenvolvimento do *Gephi* (https://github.com/gephi/gephi-plugins), existem três projetos:

- o projeto `Gephi Plugins` que funciona como base e interface com a plataforma *Gephi*
- o projeto `CGEE Insight Net 3.0` que contém o próprio código, é referenciado pelo projeto `Gephi Plugins` e referencia o projeto `InsightNetLibrary`
- o projeto `Correlation Statistics`, que fornece estatísticas adicionais.

Deve ser observado que o número de versão do *plugin* `CGEE Insight Net 3.0` deve ser o mesmo da biblioteca `InsightNetLibrary`.

Os projetos `CGEE Insight Net 3.0` e `Correlation Statistics` não podem ser compilados ou executados diretamente. Essas operações sempre devem ser realizadas a partir do projeto `Gephi Plugins`. Informações adicionais constam na página já mencionada.

9.2.3 Documentação do *CGEE Insight Net*

Esta documentação foi criada com a ferramenta “Sphinx” [8] . O código-fonte consta no repositório SVN do CGEE e pode ser editado e transformado em documentos PDF e HTML com o seguinte conjunto de ferramentas:

- Eclipse (a partir da versão Oxygen)
- Eclipse plugin ReST editor do Eclipse Marketplace
- Python, a partir da versão 3.6
- Sphinx 1.8.2
- Miktex www.miktex.org
- Perl www.perl.org

Ainda devem ser instalado o módulo Python `sphinx_bootstrap_theme` com o programa `pip`:

```
pip install sphinx_bootstrap_theme
```

A partir disso, pode ser criada uma configuração de execução no Eclipse:

1. Criar *Run configuration* no Eclipse: Usar o “make.exe” do próprio projeto `Manual InsightNet` e marcar “html” e “latexpdf” como formatos de saída
2. Compilar pelo *Run project* e verificar os arquivos gerados na pasta `build`. Destaca-se que a documentação no formato PDF consta dentro da pasta `build/latex`.

Existe ainda uma integração com a ferramenta *Gitlab* que executa todo o processo em um *container Docker* após um *commit* no repositório:

```
image: jbliesener/sphinx-doc-18-portuguese

pages:
```

[8] https://www.sphinx-doc.org

```
  script:
  - make html
  - make latexpdf
  # - make docx
  # - make text
  - mkdir public
  - mkdir public/html
  - cp -r build/html/* public/html
  - mkdir public/pdf
  - cp build/latex/*.pdf public/pdf
  # - mkdir public/docx
  # - cp build/docx/*.docx public/docx
  # - mkdir public/text
  # - cp -r build/text/* public/text
  artifacts:
    paths:
    - public
  only:
  - master
  tags:
  - sphinx
```

9.3 Uso do *CGEE Insight Net* a partir da linha de comando

9.3.1 Configuração Inicial

A linha de comando é executada a partir dos arquivos `.jar` que constam no diretório `%APPDATA%¥.gephi¥0.9.2¥dev¥modules¥`. Este caminho precisa ser informado ao Java a partir do parâmetro `-cp`. Como o caminho especificada pelo variável `%APPDATA%` geralmente contém espaços, é necessário incluir todo este caminho em aspas duplas.

O nome da classe que executa os comandos é `com.bliesener.cgee.cmdline.CommandLineParser`. Depois do nome da classe seguem os parâmetros do banco de dados:

- `-d <nome da classe do driver jdbc>` [para bancos H2 locais é `org.h2.Driver`]
- `-U <nome do usuário do banco de dados>` [para bancos H2 locais, geralmente é `cgee`]
- `-p <senha do usuário do banco de dados>` [para bancos H2 locais, geralmente é `cgee`]
- `-u <url do banco de dados>` [para bancos H2 locais, geralmente é `jdbc:h2:<caminho do banco>`]

Assim, os parâmetros básicos para executar um comando são, geralmente:

```
java -cp "%APPDATA/.gephi/0.9.2/dev/modules/*"
com.bliesener.cgee.cmdline.CommandLineParser
  -d org.h2.Driver -U cgee -p cgee -u jdbc:h2:<caminho do banco> <comando a ser
executado>
```

Para simplificar este comando, sugere-se gravar os parametros básicos em variáveis de ambiente:

```
SET CP="%APPDATA%/.gephi/0.9.2/dev/modules/*"
SET DBPARAMS=-d org.h2.Driver -U cgee -p cgee -u "jdbc:h2:<caminho do banco>"
SET CGEEPARAMS=-cp %CP% com.bliesener.cgee.cmdline.CommandLineParser %DBPARAMS%
```

A partir desta configuração, a linha de comando pode ser executada com o comando

```
java %CGEEPARAMS% <comando a ser executado>
```

9.3.2 Ajuda

O comando

```
java %CGEEPARAMS% --help
```

mostra todos os comandos e parâmetros que a ferramenta reconhece.

9.3.3 Protocolo de execução

Os seguintes parâmetros podem ser especificados ANTES do comando:

- `-l <logfile>`: Grava o protocolo de execução no arquivo especificado
- `-v <logLevel>`: Especifica o nível de protocolo. <logLevel> pode ser um dos seguintes: - OFF, - SEVERE - WARNING - INFO - CONFIG - FINE - FINER - FINEST - ALL
- `-r <reportfile>`: Grava o relatório de execução no arquivo especificado
- `--createreport`: Se o parâmetro `-r` for especificado e o arquivo referenciado já existir, os dados existentes são apagados.

9.3.4 Comandos

A ferramenta disponibiliza os seguintes comandos:

- newimport
- importbibtex
- importgeneric
- select/unselect
- newjob
- similaritysearch/tfidfsearch/keywordcooccurrence
- consolidate
- backup/restore
- shell
- batch
- statistics
- reduce
- exportgexf
- version
- query

9.3.5 `newimport`

O comando `newimport` permite a importação de currículos Lattes na ferramenta.

```
newimport      Read CVs from directory or web service
  Usage: newimport [options] Files or directories to import
    Options:
      --delete
        Delete all tables in database before starting (CAUTION!)
        Default: false
      --downloadarchivedir
        Download compressed archive of all CVs to this directory
      --filter
        Use filter expression to select input data from web service
      --info
        Information to store in the 'info' field for each imported CV
      --keyword
        Use keyword to select input data from web service
      --password
```

```
      Password to access web service
    --repair
      Repair known encoding errors
      Default: false
    --username
      User name to access web service
    --ws
      Import from WebService instead of using individual files
      Default: false
```

9.3.5.1 Parâmetros

Como parâmetro(s) obrigatório(s), o comando espera o nome do(s) diretório(s) ou arquivo(s) a serem importados. Se o nome especificado representa um diretório, todos os arquivo nele serão importados.

O identificador único usado para o currículo importado é o nome do arquivo, já que esta informação nem sempre consta nos dados XML.

`--delete`: Apagar dados antes da importação Se este parâmetro for especificado, todos os dados do banco de dados serão apagados antes da importação.

Sem este parâmetro, os dados lidos serão adicionados ao banco existente ou mesclados com ele. Para identificar se um currículo já existe no banco, é usado o identificador (nome do arquivo sem anexo .xml). Se o currículo já existir, a data da última atualização é comparada. Se o currículo importado for mais novo do que o já existente, os dados deste currículo são eliminados do banco e o arquivo é importado.

`--info`: Gravar campo de informação Este parâmetro permite a especificação de um texto que será gravado no campo “info” do banco de dados. Textos com espaços devem ser especificados entre aspas duplas

`--repair`: Pré-processar os currículos importados para eliminar caracteres inválidos.

`--ws`: Recuperar dados do *web service* Este parâmetro permite a importação dos dados via *web service* do CGEE. Mesmo especificando o parâmetro `--ws`, ainda é necessário definir um diretório de importação, **que não será usado**.

`--downloadarchivedir`: Baixar currículos para um diretório específico Se esse parâmetro for especificado, os currículos baixados pelo *web service* (usando a opção `--ws`) serão salvos no diretório especificado, em uma pasta compactada. O caminho do diretório deve estar especificado com aspas duplas, e caso este não exista, será criada uma pasta nova no caminho especificado.

`--keyword` e `--filter`: Selecionar dados do *web service* Estes parâmetros selecionam dados do *web service* por palavra-chave (`--keyword`) ou por filtro (`--filtro`) e devem ser usados em conjunto com o parâmetro `--ws`. A palavra chave ou filtro utilizado deve estar separado com aspas duplas.

`--username`: Definir nome de usuário usado para acessar o *web service*. Usado em conjunto com a opção `--ws`.

`--password`: Definir senha para o nome de usuário definido com a opção `--username`. Usado em conjunto com a opção `--ws`

9.3.5.2 Exemplos

```
JAVA %CGEEPARAMS% newimport --delete --ws --repair --info "INCT para Mudancas
Climaticas" --downloadarchivedir "C:/Users/Documents/Diretorio_CVs" --username
"user" --password "admin" --keyword “mudanças climáticas”
JAVA %CGEEPARAMS% newimport -info Demo "C:¥users¥jorg
bliesener¥documents¥cgee¥Teste100"
JAVA %CGEEPARAMS% newimport --ws --username "user" --password "password" --filter
"cpf: cpf1,cpf2,cpf3"
```

9.3.6 `importbibtex`

O comando `importbibtex` permite a importação de arquivos do formato bibtex na ferramenta:

```
importbibtex      Read Bibtex files from directory
     Usage: importbibtex [options] Files or directories to import
       Options:
            --charset
           Character set to use on import
            --delete
           Delete all tables in database before starting (CAUTION!)
           Default: false
            --forcelanguage
           Set language for all records
           Default: false
            --info
           Information to store in the 'info' field for each imported CV
            --keywords
           Type of keywords to import, (standard, author [scopus only], plus
           [WoS only]
            --language
           Language to use for records with unknown language or for all
           records (see --replacelanguage)
           Default: English
```

9.3.6.1 Parâmetros

Como parâmetro(s) obrigatório(s), o comando espera o nome do(s) diretório(s) ou arquivo(s) a serem importados. Se o nome especificado representa um diretório, todos os arquivo nele serão importados.

`--charset`: Especifica a codificação de caracteres usada nos arquivos a serem importados. Os valores sugeridos são `UTF-8` e `windows-1252`.

`--delete`: Apagar dados antes da importação Se este parâmetro for especificado, todos os dados do banco de dados serão apagados antes da importação. Sem este parâmetro, os dados lidos serão adicionados ao banco existente ou mesclados com ele.

`--info`: Gravar campo de informação Este parâmetro permite a especificação de um texto que será gravado no campo *info* do banco de dados. Textos com espaços devem ser especificados entre aspas duplas.

`--keywords`: Determina quais palavras-chave serão importadas do arquivo BibTeX. Este parâmetro pode ser especificado mais que uma vez para importar tipos de palavras-chave diversos. Valores válidos:

- `standard`: Palavras-chave padrão
- `author keywords`: Palavras-chave especificados pelo autod (apenas para arquivos

Scopus®),
- `keywords plus`: Palavras-chave adicionais (apenas para arquivos *Web of Science*)

`--forcelanguage`: Trata todas as entradas contidas nos arquivos importados como sendo de uma lingua específica.

`--language`:

- Se este parâmetro for especificado junto com `--forcelanguage`: Trata *todas* as referências bibliográficas importadas como o da linguagem determinada nessa opção.
- Se este parâmetro for especificado *sem* `--forcelanguage`: Trata as as referências bibliográficas importadas *que não especificam um idioma* como o da linguagem determinada nessa opção.

Opções:

- `English`
- `Portuguese`

9.3.6.2 Exemplo

```
importbibtex --delete "c:¥users¥jblie_000¥Documents¥CGEE¥Contrato 2016¥Bibtex com erro¥Klionsky2012445_short.bib"
```

9.3.7 `importgeneric`

O comando `importgeneric` permite a importação de dados bibliográficos genéricos, usando arquivos JSON para definir o formato dos arquivos importados (ver Seção 9.1):

```
importgeneric     Read generic bibliography files from directory, using
       format descriptor
  Usage: importgeneric [options] Files or directories to import
    Options:
      --charset
        Character set to use on import
      --delete
        Delete all tables in database before starting (CAUTION!)
        Default: false
      --descriptor
        Custom descriptor file
      --forcelanguage
        Set language for all records
        Default: false
      --info
        Information to store in the 'info' field for each imported CV
      --keywords
        Type of keywords to import, (author, database]
      --language
        Language to use for records with unknown language or for all
        records (see --replacelanguage)
        Default: English
      --saveDescriptor
        Export effective descriptor to file
      --sheet
        Sheet name or index when importing fromm Excel files. Enclose
        number in double quotes to force using sheet name instead of sheet
        index
      --type
        Generic descriptor type
```

```
       Possible Values: [Autodetect Scopus format, Comma-separated data from
Scopus,
         Tagged data from Scopus, Autodetect Web of Science format,
         Comma-separated data from Web of Science, Tab-separated data from Web
of Science,
         Tagged data from Web of Science, Generic table data (Autodetect format),
         Use external descriptor file]
```

9.3.7.1 Parâmetros

Como parâmetro(s) obrigatório(s), o comando espera o nome do(s) diretório(s) ou arquivo(s) a serem importados. Se o nome especificado representa um diretório, todos os arquivo nele serão importados.

`--charset`: Especifica a codificação de caracteres usada nos arquivos a serem importados. Os valores sugeridos são `UTF-8` e `windows-1252`.

`--delete`: Apagar dados antes da importação Se este parâmetro for especificado, todos os dados do banco de dados serão apagados antes da importação. Sem este parâmetro, os dados lidos serão adicionados ao banco existente ou mesclados com ele.

`--info`: Gravar campo de informação Este parâmetro permite a especificação de um texto que será gravado no campo *info* do banco de dados. Textos com espaços devem ser especificados entre aspas duplas.

`--sheet`: Especificar nome ou número da aba em planilhas Excel® Para importações de planilhas Excel®, este parâmetro permite definir o nome ou o número da aba em que constam os dados a serem importados.

`--type`: Formato de dados. O parametro `--type` é obrigatório e determina o formato dos dados de entrada. Os seguintes valores são aceitos e devem ser especificados em aspas duplas:

- `Autodetect Scopus format`
- `Comma-separated data from Scopus`
- `Tagged data from Scopus`
- `Autodetect Web of Science format`
- `Comma-separated data from Web of Science`
- `Tab-separated data from Web of Science`
- `Tagged data from Web of Science`
- `Generic table data (Autodetect format)`
- `Use external descriptor file`

`--descriptor`: Nome do arquivo que especifica o formato do arquivo a ser importado, caso for definido `type "Use external descriptor file"`

`--saveDescriptor`: Exportar o arquivo de especificação do formato para o arquivo especificado

`--keywords`: Determina quais palavras-chave serão importadas do arquivo BibTeX. Este parâmetro pode ser especificado mais que uma vez para importar tipos de palavras-chave diversos. Valores válidos:

- `standard`: Palavras-chave padrão
- `author keywords`: Palavras-chave especificados pelo autod (apenas para arquivos *Scopus*®),
- `keywords plus`: Palavras-chave adicionais (apenas para arquivos *Web of Science*)

`--forcelanguage`: Trata todas as entradas contidas nos arquivos importados como sendo de uma lingua específica.

`--language`:

- Se este parâmetro for especificado junto com `--forcelanguage`: Trata *todas* as referências bibliográficas importadas como o da linguagem determinada nessa opção.
- Se este parâmetro for especificado *sem* `--forcelanguage`: Trata as as referências bibliográficas importadas *que não especificam um idioma* como o da linguagem determinada nessa opção.

Opções:

- `English`
- `Portuguese`

9.3.8 `select`/`unselect`

Os comandos `select` e `unselect` definem quais contribuições científicas dos pesquisadores farão parte da análise de coautoria e de similaridade, e se o leitor de contribuições `SelectedContributionReader` será usado.

```
select     Add contributions to/Remove contributions from subset to
       process
  Usage: select [options]
    Options:
      --after
        use with --title: select/unselect only contributions within this
        many years after the researcher has obtained the specified title
        Default: 0
      --all
        add/remove all contributions
        Default: false
      --defaultContributionSelect
        use with --title: select/unselect all contributions that do not
        have a publication date
        Default: false
      --defaultResearcherSelect
        use with --title: select/unselect all contributions for reseachers
        that don't have the specified title
        Default: false
      --title
        select/unselect contributions after this title
      --where
        filter expression

unselect     Add contributions to/Remove contributions from subset to
       process
  Usage: unselect [options]
    Options:
      --after
        use with --title: select/unselect only contributions within this
        many years after the researcher has obtained the specified title
        Default: 0
      --all
        add/remove all contributions
        Default: false
      --defaultContributionSelect
        use with --title: select/unselect all contributions that do not
        have a publication date
        Default: false
      --defaultResearcherSelect
```

```
       use with --title: select/unselect all contributions for reseachers
       that don't have the specified title
       Default: false
     --title
       select/unselect contributions after this title
     --where
       filter expression
```

9.3.8.1 Parâmetros

`--where`: Permite a especificação de uma cláusula de escopo. Esta cláusula segue a sintaxe do JPA (ver http://www.objectdb.com/java/jpa/query/jpql/where), sendo que o comando completo começa com `SELECT bc FROM BASECONTRIBUTION ...`.

Os campos e relacionamentos devem seguir o modelo de banco de dados da ferramenta (ver documentação e diagrama de entidades e relacionamentos).

`--all`: Seleciona/Desseleciona todas as contribuições.

`--title`: Seleciona/Desseleciona contribuições depois de um título específico (Ex: Mestrado, Doutorado). O programa só levará em consideração as publicações feitas depois do título especificado.

`--after`: Seleciona/Desseleciona contribuições contidas no período definido de anos depois do título. Deve ser usada com a opção `--title`.

`--defaultContributionSelect`: Seleiona/Desseleciona todas as publicações sem data especificada. Deve ser usada com a opção `--title`.

9.3.8.2 Exemplos

```
java %CGEEPARAMS% select --where "TYPE(bc)=Artigo"
java %CGEEPARAMS% select --where "TYPE(bc)=CapituloLivro AND bc.ano>=1990 AND
bc.ano<=2015"
java %CGEEPARAMS% select --where "TYPE(bc)=TrabalhoEmEventos AND TREAT(bc as
TrabalhoEmEventos).workType=com.bliesener.cgee.entities.TrabalhoEmEventos.Wor
kType.COMPLETO AND bc.ano>=1965 AND bc.ano<=2015"
```

9.3.9 `newjob`

O comando `newjob` prepara o banco de dados para a execução de uma pesquisa de similaridade e define o escopo da mesma.

9.3.9.1 Parâmetros

Os parâmetros se dividem em três grupos:

- Parâmetros que constam no diálogo da pesquisa por similaridade da ferramenta:
  - `--casesensitive`
  - `--partitionyear`
  - `--dolevenshtein`
  - `--minsimilarity` e `--minsimilaritylevenshtein` (sinónomos)
  - `--dotfidf`
  - `--content`
  - `--dostemming`
  - `--usestopwords`
  - `--minsimilaritytfidf`
  - `--sparse`
  - `--dokeywordcooccurrence`
  - `--minkeywordcooccurence`

    - `--selectedcontributionsonly`
- Parâmetros que constam na tela principal de configurações da ferramenta:
    - `--blocksize`
    - `--useoldlimitformula`
    - `--usealternativelevenshtein`
    - `--poolsametitle`
    - `--usebktreelimit`
    - `--tflog`
    - `--tfidfpercentile`
    - `--roundtfidf`
- Parâmetros de seleção de escopo - Ver próximo parágrafo:
    - `--contributionreader`
    - `--entitytype`
    - `--whereclause`
    - `--type`
    - `--update`

9.3.9.2 Seleção de Escopo

Durante a fase de desenvolvimento da ferramenta, o escopo de análise de similaridade variou e foi necessária uma forma flexível para definí-lo. Hoje existem quatro métodos para especificar quais contribuições serão incluidas na pesquisa de similaridade:

1. Todas as contribuições: Nenhum dos parâmetros `--contributionreader`, `--entitytype`, `--whereclause` será especificado.
2. Contribuições de apenas um tipo específico. O parâmetro `--entitytype` especifica este tipo:
    1. `java %CGEEPARAMS% newjob --dolevenshtein --entitytype Artigo`
    2. `java %CGEEPARAMS% newjob --dolevenshtein --entitytype CapituloLivro`
    3. `java %CGEEPARAMS% newjob --dolevenshtein --entitytype TrabalhoEmEventos`
3. Contribuições que atendem uma cláusula `WHERE` específico, de acordo com a sintaxe do JPA (veja http://www.objectdb.com/java/jpa/query/jpql/where):
    1. `java %CGEEPARAMS% newjob --dolevenshtein --whereclause "TYPE(bc)=CapituloLivro AND bc.ano>=1990 AND bc.ano<=2015"`
4. Para escopos mais complexos, como aquele usado frequentemente no CGEE (todos os Artigos, todos os capítulos de livros e apenas trabalhos completos em eventos, pode ser especificada uma classe java específica que fornece estas contribuições. Para isso, existe a classe `com.bliesener.cgee.similarity.SelectedContributionReader`, que retorna todas as contribuições selecionadas com os comandos `select`/`unselect`:
    1. `java %CGEEPARAMS% newjob --dolevenshtein --contributionreader com.bliesener.cgee.similarity.SelectedContributionReader`

9.3.9.3 Exemplos

```
JAVA %CGEEPARAMS% newjob --minsimilaritytfidf 0.05 --partitionyear
JAVA %CGEEPARAMS% newjob --contributionreader SelectedContributionReader
--dolevenshtein false --dostemming false --minsimilaritytfidf 0.15
```

9.3.10 `similaritysearch`/`tfidfsearch`/`keywordcooccurrence`

O comando `similaritysearch` executa a pesquisa por co-autoria definida com `newjob`. O comando `tfidfsearch` executa a pesquisa definina com newjob, realizando a busca por similaridade semântica. O comando `keywordcooccurrence` cria redes de co-ocorrências de palavras-chaves.

```
similaritysearch     Start similarity search on database
  Usage: similaritysearch [options]
    Options:
          --threads
         Set number of concurrent threads
         Default: 1

tfidfsearch     Start tf.idf similarity search on database
  Usage: tfidfsearch [options]
    Options:
          --threads
         Set number of concurrent threads
         Default: 1

keywordcooccurrence     Create network of keyword co-occurrences
  Usage: keywordcooccurrence [options]
    Options:
      --threads
        Set number of concurrent threads
        Default: 1
```

9.3.10.1 Parâmetros

`--threads`: Define a quantidade de execuções em paralelo e corresponde ao respectivo parâmetro definido na tela de opções da ferramenta visual. Recomenda-se um valor que fica abaixo da quantidade de núcleos do computador usado.

9.3.10.2 Exemplos

```
JAVA %CGEEPARAMS% tfidfsearch --threads 7
JAVA %CGEEPARAMS% similaritysearch --threads 7
JAVA %CGEEPARAMS% keywordcooccurrence --threads 7
```

9.3.11 `consolidate`

O comando `consolidate` processa os resultados da pesquisa **por coautorias** e gera as arestas do grafo no banco de dados.

```
consolidate     Consolidate similarity search results
   Usage: consolidate [options]
     Options:
          --checksymmetry
         Check graph symmmetry
         Default: false
          --checksymmetryonly
         Check graph symmmetry only, do not consolidate
         Default: false
```

9.3.11.1 Parâmetros

Os parâmetros servem apenas para a depuração do procedimento.

9.3.11.2 Exemplos

```
JAVA %CGEEPARAMS% consolidate
```

9.3.12 `backup`/`restore`

Os comandos `backup` e `restore` permitem a gravação do banco de dados em arquivos

externos e a sua recuperação.

```
backup      Backup internal database
   Usage: backup [options]
     Options:
       --filename
          Output file name
       -z, --zip
          Zip compress output file
          Default: false

 restore      Restore internal database from backup
   Usage: restore [options]
     Options:
           --filename
          Input file name
       -z, --zip
          Input file is zip compressed
          Default: false
```

9.3.12.1 Parâmetros

`--filename`: Nome do arquivo a ser carregado/salvo

`--zip`: Define se o arquivo de input/output é uma pasta comprimida ou não.

9.3.12.2 Exemplos

```
JAVA %CGEEPARAMS% backup --filename “Test”
JAVA %CGEEPARAMS% restore --filename “Test” --zip
```

9.3.13 `shell`

O comando `shell` abre um interpretador de linha de comando que permite a digitação de comandos interativos.

```
shell      Enter interactive shell
  Usage: shell [options]
```

9.3.14 `batch`

O comando `batch` permite a execução de vários comandos gravados em arquivo.

```
batch      Run commands from batch file
  Usage: batch [options] Batch file to run
```

9.3.14.1 Parâmetros

Como parâmetro do comando deve ser especificado o nome do arquivo que contêm os comandos a serem executados.

9.3.14.2 Exemplo

```
JAVA %CGEEPARAMS% -v FINE -l c:¥temp¥cgee.log batch c:¥temp¥cgeeCommands.txt
```

9.3.15 `reduce`

O comando `reduce` remove todos os nós do banco de dados exceto aqueles selecionados.

```
reduce      Remove all but a specific set of nodes from network
  Usage: reduce [options] List of remaining ids in network
```

9.3.15.1 Parâmetros

Como parâmetro do comando devem ser especificados os ids dos nós que deverão permanecer na rede.

9.3.15.2 Exemplo

```
JAVA %CGEEPARAMS% reduce “id1, id2, id3”
```

9.3.16 statistics

O comando `statistics` mostra as informações gerais sobre a base.

```
statistics      Provide database statistics
   Usage: statistics [options]
```

9.3.17 exportgexf

O comando `exportgexf` exporta a rede contida no banco de dados como um arquivo do tipo `.gexf`.

```
exportgexf      Export Graph to GEXF file
     Usage: exportgexf [options] Name of file to export
       Options:
            --withinfo
           Export the 'info' field for each Pesquisador
           Default: false
```

9.3.17.1 Parâmetros

Como parâmetro obrigatório do comando deve ser fornecido o nome do arquivo de output.

`--withinfo`: Define se o campo ‘info’ dos nós também será exportado.

9.3.17.2 Exemplo

```
JAVA %CGEEPARAMS% exportgexf  --withinfo “Test.gexf”
```

9.3.18 version

O comando `version` mostra a versão atual da biblioteca.

```
version      Print library version
   Usage: version [options]
```

9.3.19 query

O comando query permite realizar pesquisas JPA diretamente na base de dados do *CGEE Insight Net*. Esta funcionalidade serve, principalmente, para depurar o módulo.

9.3.19.1 Parâmetros

O parâmetro obrigatório é a pesquisa a ser realizada.

`--limit`: Limita a quantidade de resultados.

9.3.19.2 Exemplo

```
java %cgeeparams% query "SELECT COUNT(bc) FROM BaseContribution bc"
```

9.3.20 Fluxo de Trabalho

Geralmente, um fluxo de criação de redes segue a seguinte sequência de comandos:

1. Importação dos CVs/Arquivos Bibtex/Referências bibliográficas - comando newimport

ou importbibtex ou importgeneric

2. Seleção das publicações desejadas - comandos select/unselect OBS: Nessa etapa são escolhidos os tipos de publicação e o período de publicações desejado
3. Determinação dos parâmetros de busca - comando newjob OBS: Nessa etapa são definidas a maior parte das opções contidas na tela usada para rodar a rede.
4. Execução da busca por coautorias e/ou similaridade semântica - comandos similaritysearch/tfidfsearch/keywordcooccurrence
5. Consolidação dos resultados - comando consolidate
6. Exportação dos resultados - comandos exportgexf e/ou backup/restore

Segue um exemplo de um fluxo de trabalho de análise de rede de pesquisadores.

```
java %CGEEPARAMS% newimport --delete --info "Test100" "c:¥users¥jorg
bliesener¥documents¥cgee¥teste100"
java %CGEEPARAMS% select --where "TYPE(bc)=Artigo"
java %CGEEPARAMS% select --where "TYPE(bc)=CapituloLivro"
java %CGEEPARAMS% select --where "TYPE(bc)=TrabalhoEmEventos AND TREAT(bc as
TrabalhoEmEventos).workType=com.bliesener.cgee.entities.TrabalhoEmEventos.Wor
kType.COMPLETO"
java %CGEEPARAMS% newjob --dolevenshtein --partitionyear --contributionreader
com.bliesener.cgee.similarity.SelectedContributionReader
java %CGEEPARAMS% similaritysearch --threads 7
java %CGEEPARAMS% consolidate
```

Depois da execução destes comandos, a rede consta no banco de dados e pode ser extraída a partir de comandos SQL.

O mesmo fluxo pode ser realizado em arquivo batch, de acordo com o exemplo em seguida. A vantagem é que a conexão com o banco de dados precisa ser feita apenas uma única vez. Neste exemplo, o arquivo terá o nome `cmdline.batch`.

```
# importar os curriculos
import --delete --info "Test100" "c:¥users¥jorg
bliesener¥documents¥cgee¥teste100"
# selecionar todos os Artigos, todos os capítulos de livros e apenas os trabalhos
em eventos COMPLETOS
select --where "TYPE(bc)=Artigo"
select --where "TYPE(bc)=CapituloLivro"
select --where "TYPE(bc)=TrabalhoEmEventos AND TREAT(bc as
TrabalhoEmEventos).workType=com.bliesener.cgee.entities.TrabalhoEmEventos.Wor
kType.COMPLETO"
# realizar pesquisa levenshtein e consolidar os resultados
newjob --dolevenshtein --partitionyear --contributionreader
com.bliesener.cgee.similarity.SelectedContributionReader
similaritysearch --threads 7
consolidate
```

Para executar o fluxo e gravar os resultados com protocolo fino em arquivo de protocolo `cgeeLog.log` pode ser usado o seguinte comando:

```
java %CGEEPARAMS% -l cgeeLog.log -v FINE batch cmdline.txt
```

No exemplo abaixo, é realizada uma extração de currículos usando um filtro de CPF através do webservice. Depois, são selecionadas as publicações do tipo Artigo Completo, Capítulo de Livro, e Partipações em Eventos Completas, no período de 1999 a 2007. A rede é rodada com similaridade semântica mínima de 0.05, e tanto a busca por coautorias quanto por similaridade semântica são realizadas. Depois, a rede gerada é exportada para

um arquivo `.gexf`:

```
JAVA %CGEEPARAMS% newimport --delete --ws --repair --info "Física Quântica"
--downloadarchivedir "C:/Users/Documents/minhas_redes" --username "user"
--password "password" --filter "cpf:cpf1, cpf2, cpf3;"
JAVA %CGEEPARAMS% unselect --all
JAVA %CGEEPARAMS% select --where "TYPE(bc)=Artigo AND TREAT(bc as
Artigo).workType=com.bliesener.cgee.entities.Artigo.WorkType.COMPLETO AND
bc.ano>=1999 AND bc.ano<=2007"
JAVA %CGEEPARAMS% select --where "TYPE(bc)=CapituloLivro AND bc.ano>=1999 AND
bc.ano<=2007"
JAVA %CGEEPARAMS% select --where "TYPE(bc)=TrabalhoEmEventos AND TREAT(bc as
TrabalhoEmEventos).workType=com.bliesener.cgee.entities.TrabalhoEmEventos.Wor
kType.COMPLETO AND bc.ano>=1999 AND bc.ano<=2007"
JAVA %CGEEPARAMS% newjob --minsimilaritytfidf 0.05 --partitionyear
JAVA %CGEEPARAMS% similaritysearch --threads 7
JAVA %CGEEPARAMS% tfidfsearch --threads 7
JAVA %CGEEPARAMS% consolidate
JAVA %CGEEPARAMS% exportgexf --withinfo "C:/Users /Documents/Minhas
Redes/rede_antes.gexf"
```

9.4 Troca de banco de dados da base Lattes

A alteração no final do ano do banco de dados do nosso espelho Lattes do DB2 para PostGreSQL, juntamente com o extenso trabalho de limpeza nos dados da base Lattes local realizado pela TI e atualizações na ferramenta de gerenciamento da API do extrator Lattes determinaram a necessidade de adaptações na versão 3.2.6, que estava testada, validada e pronta para entrega para os usuários. Apesar da versão de teste já taer sido disponibilida para usuários internos interessados, foi decidido que a entrega para os demais usuários seria adiada até a incorporação das alterações necessárias. Assim, para todos os efeitos práticos, a versão 3.2.7, que será basicamente apenas a 3.2.6 com as alterações, quando estiverem terminadas, terá as mesmas características registradas neste relatório.