Centro de Gestão e Estudos Estratégicos
Ciência, Tecnologia e Inovação

Projeto Temático:
Apoio técnico ao CNPq para o desenvolvimento e aprimoramento de metodologias de planejamento estratégico

Relatório da capacitação de analistas do CNPq

Projeto Temático:
Apoio técnico ao CNPq para o desenvolvimento e aprimoramento de metodologias de planejamento estratégico

Relatório da capacitação de analistas do CNPq

Brasília, DF
Dezembro, 2017

Centro de Gestão e Estudos Estratégicos

Relatório da capacitação de analistas do CNPq. Projeto Temático: Apoio técnico ao CNPq para o desenvolvimento e aprimoramento de metodologias de planejamento estratégico Brasília: Centro de Gestão e Estudos Estratégicos, 2017.

47 p. : il.

1. Análise de dados. 2. Conceitos. 3. Boas práticas. 4. Metodologias. I. CGEE. II. Título.

Centro de Gestão e Estudos Estratégicos - CGEE
SCS Qd 9, Lote C, Torre C
Ed. Parque Cidade Corporate - salas 401 a 405
70308-200 - Brasília, DF
Telefone: (61) 3424.9600
Fax. (61) 3424 9659
http://www.cgee.org.br

Este documento é parte integrante das atividades desenvolvidas no âmbito do 2º Contrato de Gestão CGEE – 11º Termo Aditivo/Atividade - Desenvolvimento de Competências e Ferramentas em Prospecção, Avaliação Estratégica, Gestão da Informação e do Conhecimento/Projeto: Exploração de Dados e Visualização de Informação - 56.2.81.04/ MCTI/2016.

Projeto Temático:
Apoio técnico ao CNPq para o desenvolvimento e aprimoramento de metodologias de planejamento estratégico

Relatório da capacitação de analistas do CNPq

Supervisão
Marcio de Miranda Santos

Coordenação
Jackson Maia

Equipe técnica
Cristiano Alves da Silva (Estagiário)
Eduardo Moresi
Gabriel Fritz Sluzala (Estagiário)
Rogério da Silva Castro
Sofia Daher

Sumário

CAPÍTULO 1 - Introdução 6
1.1 Contexto e Visão Geral 6

CAPÍTULO 2 - Motivações e antecedentes do Projeto 8
2.1 Motivações 8
2.2 Prova de conceito – avaliação preliminar de resultados do Programa SISBIOTA Brasil 9
2.3 Plano de trabalho básico para avaliações de resultados 14

CAPÍTULO 3 - Evento de capacitação de analistas do CNPq 16
3.1 Programa da capacitação 16
3.2 Contexto metodológico da capacitação 17
3.3 Benefícios esperados da capacitação 21

Anexo A Slides selecionados das apresentações da capacitação 22

CAPÍTULO 1 - Introdução

1.1 Contexto e Visão Geral

O vertiginoso desenvolvimento científico-tecnológico vivenciado nas últimas décadas tende a ser ainda mais acelerado nas décadas seguintes, de tal forma que se esperam cenários radicalmente diferentes da conjuntura atual, com reflexos profundos no cotidiano da sociedade. Com isso, novos quadros políticos, econômicos e sociais poderão se apresentar de modo diverso ao atual para os extratos socioeconômicos do País.

Também é crescente a geração e digitalização da informação, particularmente a de origem científico-tecnológica, assim como a disponibilização de grandes massas de dados oriundos desses setores. Esses fatos apresentam desafios e oportunidades para a atuação de agências de fomento, que precisam estar preparadas para atuar nesse novo cenário. Nesse contexto, o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) buscou a competência do CGEE para atuar no apoio à elaboração de cenários prospectivos e no desenvolvimento de métodos e ferramentas em inteligência em CT&I capazes de lidar com grandes volumes de dados oriundos de fontes distintas, de forma a incorporá-los às ações e às atividades executadas pelo Conselho.

O Projeto *Apoio Técnico ao CNPq na Utilização de Métodos e Ferramentas Modernas de Inteligência em CTI* tem como principal objetivo a Incorporação de métodos e ferramentas desenvolvidos pelo CGEE na construção de metodologias aplicadas à elaboração de cenários prospectivos de desenvolvimento institucional e avaliação de programas executados pelo CNPq. Entre suas primeiras ações, foi ministrado um curso básico nas ferramentas de inteligência em CTI empregadas pelo CGEE a analistas no CNPq, em dezembro de 2017, com a finalidade de facilitar as fases subsequentes das atividades do Projeto.

O Capítulo 2 deste relatório descreve as motivações e antecedentes para o Projeto, além dos principais resultados de uma prova de conceito para

uma avaliação de resultados do Programa SISBIOTA Brasil. A prova de conceito também foi útil para o levantamento de necessidades para o curso de capacitação básica para analistas do CNPq, ministrado em 18 e 19 de dezembro de 2017 e descrito no Capítulo 3, junto com algumas conclusões. O relatório ainda contém um Apêndice, para registro, com os principais *slides* apresentados na capacitação.

.

CAPÍTULO 2 - Motivações e antecedentes do Projeto

2.1 Motivações

Nos últimos anos, o CGEE tem concentrado esforços no desenvolvimento de métodos e ferramentas capazes de abordar o desafio de transformar grandes massas de dados de CT&I em informação relevante para apoiar tomadas de decisão com base em evidências. Gerenciando milhares de projetos por ano e com um papel central no fomento à CT&I do país, o CNPq apresenta demandas potenciais para o uso das ferramentas desenvolvidas pelo Centro.

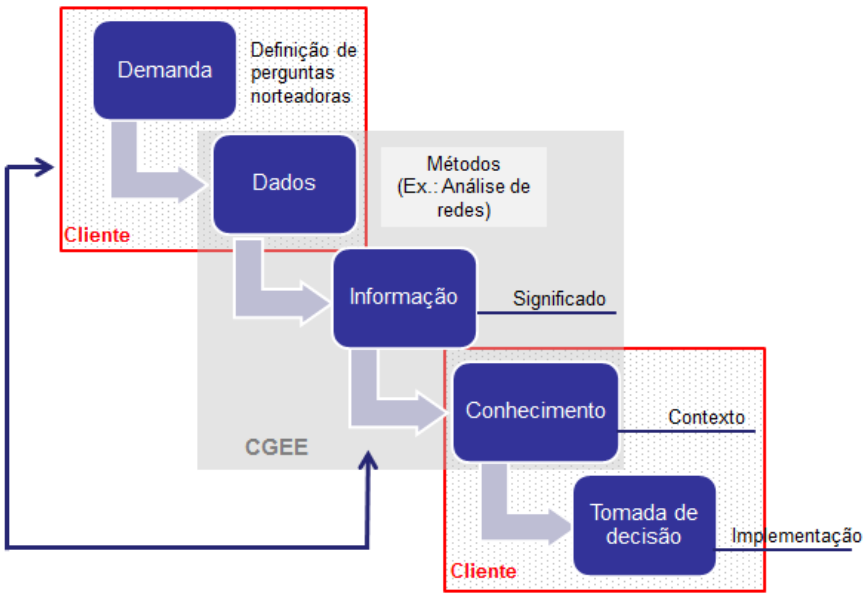
Como forma explorar a aplicabilidade dos métodos do CGEE em temas reais de interesse do CNPq, foram realizadas análises de alguns dos dados disponíveis do Programa SISBIOTA Brasil no primeiro semestre de 2017. A partir dos resultados dessa atividade, foi possível levantar necessidades mais imediatas do curso básico de capacitação para analistas do CNPq, realizado no final do ano, que iniciou o projeto propriamente dito. Esta prova de conceito também será útil para o planejamento das fases subsequentes do Projeto.

Além da capacitação, o Projeto prevê para 2018 o planejamento e a execução de avaliações de resultados de: a) um programa do “balcão” de demanda espontânea, o Programa de Ciências Ambientais, b) o aprofundamento das análises do Programa SISBIOTA Brasil, como exemplo de aplicação para programas temáticos de demanda induzida, c) uma aplicação de métodos de análise de resultados no que diz respeito a inovação do Programa Institutos Nacionais de Ciência e Tecnologia (INCT) e d) um trabalho de elaboração de cenários prospectivos para atuação futura do Conselho, como subsídio fundamental para a continuidade do Plano Estratégico Institucional. O Plano Estratégico do CNPq foi elaborado anos atrás também com o apoio do CGEE.

2.2 Prova de conceito – avaliação preliminar de resultados do Programa SISBIOTA Brasil

De acordo com seu Documento Base, o Programa SISBIOTA Brasil tem como objetivo " Ampliar a competência técnico-científica e a abrangência temática e geográfica das pesquisas em biodiversidade no Brasil". Para este proposito, o SISBIOTA aprovou 39 redes de pesquisa em todo o Território Nacional, com financiamento conjunto do CNPq, do Fundo Nacional de Desenvolvimento Científico e Tecnológico, do Ministério do Meio Ambiente, da Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior e de 13 Fundações Estaduais de Amparo à Pesquisa. O valor total financiado foi de, aproximadamente, R$ 44 milhões entre 2011 e 2015.

O ponto de partida da prova de conceito foi levantar as necessidades de análise a partir de um processo simples de tomada de decisões baseada em evidências mostrado na figura abaixo



Da elicitação de requisitos inicial foram obtidas as seguintes perguntas norteadoras:

1. Os pesquisadores do SISBIOTA BRASIL passaram a ter maior produção científica conjunta após a integração ao programa?
2. Os bolsistas e alunos do SISBIOTA BRASIL deram continuidade

em sua formação e/ou adquiriram vínculo empregatício, ou funcional, após a integração ao programa?

3. Houve a realização de pesquisas interdisciplinares, entendidas como abrangendo a integração de distintas áreas do conhecimento em torno de um problema de pesquisa?
4. Houve a formação de recursos humanos de modo interdisciplinar, abrangendo a participação de alunos e bolsistas na produção científica de distintas áreas do conhecimento em torno de um problema de pesquisa?

A partir de uma análise de tempo disponível e de disponibilidade facilitada a bases de dados, bem como o atendimento ao principal objetivo do Programa, a pergunta norteadora 1 foi escolhida para ser abordada ao longo da prova de conceito.

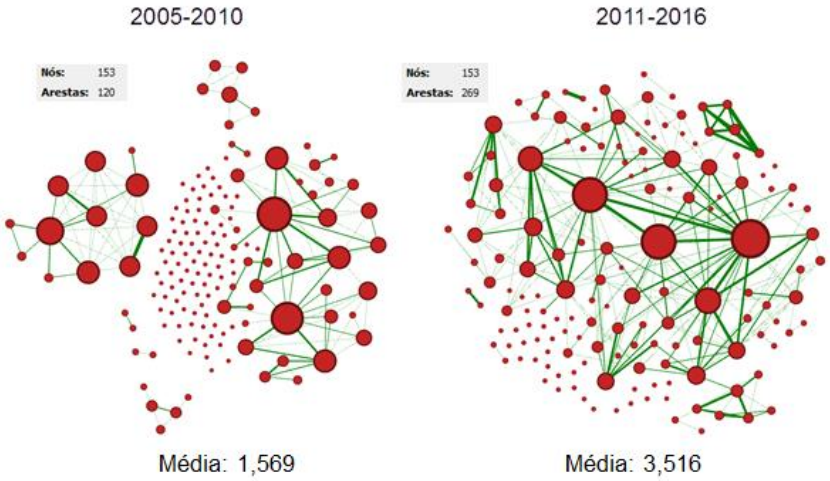
Como indicador de produção científica conjunta, optou-se por extrair dados de evolução de coautorias entre os pesquisadores das 39 redes de pesquisa. O universo de pesquisadores compreendeu os membros de projeto descritos nas propostas originais e bolsistas do CNPq e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) que se agregaram posteriormente aos projetos. Como resultado do levantamento desse universo, foram encontrados 2.275 pesquisadores, sendo que foram descartados dados que não constassem na base de currículos Lattes.

O método escolhido para a análise exploratória dos dados foi o de análise de redes complexas, com várias de suas métricas implementadas para uso em dados de currículos Lattes na ferramenta *Insight Net*, desenvolvida no CGEE para uso conjunto com a plataforma de visualização de redes *Gephi*. No contexto de redes adotado, cada nó correspondeu a um currículo, entendido aqui como o conteúdo de publicações de um pesquisador, e as arestas identificavam relações de coautoria entre pares de pesquisadores, sendo que o número de tais coautorias determinaram os pesos das arestas. Como critério de contagem de coautorias usou-se o número de arestas (ou “grau”, no jargão da área de análise de redes), como

forma de evitar duplicidades inerentes a diferentes critérios de contagem, tais como número de coautores ou número de publicações em coautoria. Para efeito da determinação de uma coautoria, considerou-se apenas a produção presumivelmente revisada por pares: artigos de periódicos, artigos completos em congressos e capítulos de livros.

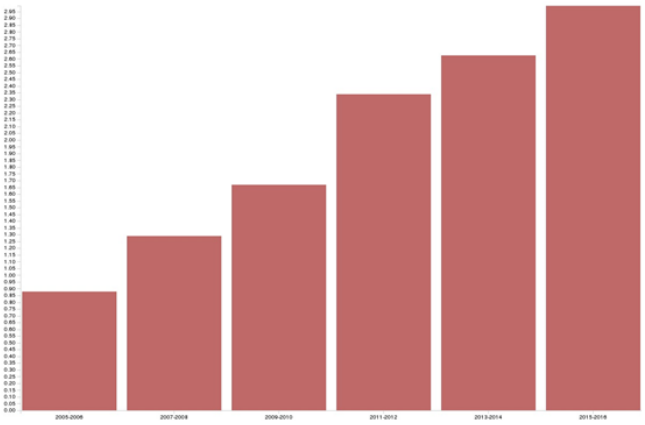
Para cada projeto foram geradas redes de coautoria, considerando a produção dos seus membros entre 2005 e 2010 e entre 2011 e 2016. Os dois intervalos foram escolhidos para efeito de comparação entre as produções publicadas desde o lançamento do Edital do SISBIOTA Brasil até 2016 e um intervalo simétrico anterior.

Tomando como exemplo o Projeto “Diversidade de Campos Sulinos”, pode-se notar um expressivo aumento da média de coautorias (vide figura)

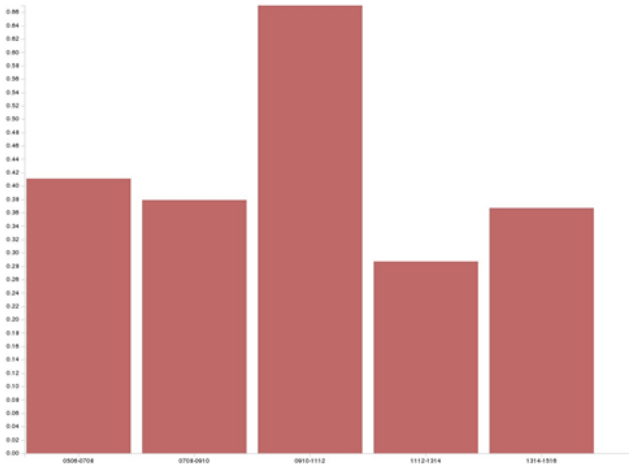


Para o conjunto completo de 39 projetos, houve uma variação do grau médio de 1,759 para 4,885 nos períodos 2005-2010 e 2011-2016, respectivamente.

Apesar de promissor, esse resultado não pode ser considerado como devido isoladamente ao Programa SISBIOTA. Dividindo-se o período completo 2005-2016 em biênios, nota-se que existe um crescimento natural no número de coautorias mesmo antes do início do Programa SSBIOTA Brasil, como se pode notar no histograma abaixo:



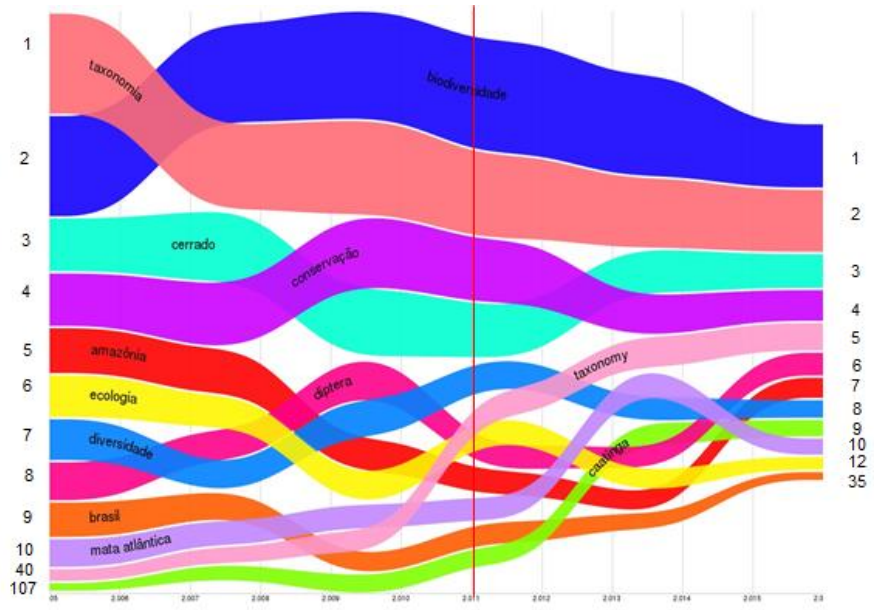
Para isolar os efeitos desse crescimento “basal”, foi produzido o histograma das diferenças entre dois biênios ao longo dos 12 anos considerados, que mostra claramente um aumento de taxa de crescimento na diferença entre o biênio 2009-2010 e o biênio 2011-2012, imediatamente antes e depois do lançamento do Programa:



A análise de causalidade, necessária em um estudo mais completo, era fora do escopo da prova de conceito, mas esse resultado sugere que o lançamento do Programa SISBIOTA Brasil teve um efeito positivo, induzido o aumento da formação de redes de pesquisa.

Para testar os possíveis efeitos do Programa na produção científica relacionada a diferentes temas de pesquisa, foram também levantadas curvas de crescimento de palavras-chave empregadas nos artigos. Na figura abaixo, são mostradas as frequências de palavras-chave por ano,

considerando apenas as evoluções das 10 mais frequentes no início do período de estudo e as 10 mais frequentes em 2016:



Algumas observações devem ser feitas acerca da figura acima:

1. Não foram excluídas multiplicidades de palavras-chave obtidas de artigos em coautoria, portanto existe um possível viés de frequência para temas que tenham artigos com muitos coautores,
2. Foi deliberadamente evitado juntar palavras-chave de mesmo significado em línguas diferentes, para evidenciar diferenças entre publicações nacionais e internacionais,
3. Notam-se crescimentos claros nas palavras-chave “mata atlântica”, “caatinga” e “taxonomy”, sendo que as duas primeiras crescem após o lançamento do edital e a última não pode ter seu crescimento considerado como correlato ao lançamento do edital,
4. De acordo com os analistas do CNPq, o crescimento da palavra-chave “taxonomy” parece estar relacionado ao lançamento de um programa específico para esse tema chamado PROTAX.

Também foram realizados testes para tentar identificar se redes que receberam financiamentos anteriores apresentavam resultados diferentes de evolução de coautorias com relação a redes que foram criadas a partir do

lançamento do Programa. Em uma primeira análise, não parece ter havido diferença.

Conclusões

De modo geral, todos os projetos tiveram aumento de coautoria média, sendo que o projeto com menor crescimento apresentou aumento de 30% entre os períodos selecionados. O projeto de maior crescimento apresentou aumento de mais de 1100% no número médio de coautorias. Embora o último número impressione, deve-se ressaltar que se tratava de um grupo que foi criado para concorrer ao Edital que não tinha muitas coautorias anteriormente.

O aumento de frequência da palavra-chave “taxonomy” antes do lançamento do Programa SISBIOTA Brasil deixa claro que esse tipo análise não pode prescindir de dados de outros programas com atuações temáticas similares, se o objetivo for isolar impactos de um dado programa temático.

O estudo preliminar mostrou caminhos promissores para análises futuras, identificou bases de dados e ferramentas relevantes para avaliações de resultados de programas temáticos, identificou melhorias que poderiam ser implementados nos futuros formulários de submissão de propostas do CNPq para enriquecer os dados disponíveis. Também como consequência da prova de conceito, foi possível elaborar um plano de trabalho e cronograma básicos a serem adaptados a casos semelhantes, apresentado na próxima seção.

2.3 Plano de trabalho básico para avaliações de resultados

A partir dos tempos despendidos pelas duas equipes de trabalho (CGEE e CNPq), a identificação de escopos típicos para a realização de avaliações de resultados no CNPq e da identificação de necessidades de capacitação da equipe envolvida na prova de conceito, foi elaborado em conjunto o seguinte plano de trabalho básico, incluindo prazos, que, presume-se, pode ser proposto e adaptado para iniciativas semelhantes em outros contextos no CNPq:

1. Oficina sobre ferramentas bibliométricas de análises de dados científicos (20 dias após o início das atividades de elicitação e refinamento de perguntas norteadoras)
2. Definição do processo específico de análise de dados para o programa e elaboração de plano de atividades (30 dias após o início)
3. Extração, tratamento e limpeza dos dados (75 dias após o início)
4. Análise exploratória dos dados (90 dias após o início)
5. Consolidação da metodologia específica de avaliação do programa (110 dias após o início)
6. Elaboração de plano de comunicação dos resultados (120 dias após o início)
7. Documento descritivo dos resultados obtidos (140 dias após o início)

Note-se que, após a primeira oficina, cada bloco de análise (passos 2-7) em princípio pode ser realizado em 120 dias e eventuais processos de monitoramento, desde que empregada a mesma metodologia, podem ser realizados em bem menos tempo. Além das metas e cronograma propostos acima, a prova de conceito da análise de resultados do SISBIOTA Brasil também foi importante para elaboração do evento de capacitação descrito no próximo capítulo.

CAPÍTULO 3 - Evento de capacitação de analistas do CNPq

3.1 Programa da capacitação

A capacitação para analistas do CNPq foi realizada nos dias 18 e 19 de dezembro de 2017, para 9 analistas e para o Diretor de Cooperação Institucional do CNPq, juntamente com 4 participantes do CGEE. O objetivo de reunir as duas equipes foi de estabelecer um vocabulário e percepção comuns para a atuação em cooperação nas fases subsequentes do Projeto "Apoio técnico ao CNPq para o desenvolvimento e aprimoramento de metodologias de planejamento estratégico". O CNPq selecionou analistas relacionados às três áreas do Projeto.

O programa do evento de capacitação cobriu os principais conceitos, métodos e ferramentas pertinentes à tomada de decisões com base em evidências e empregados no CGEE, conforme descrito abaixo:

18/12

9h30 -10h – Contexto: análise de dados para decisões baseadas em evidências

10h – 10h30 – O processo de análise de dados em CTI

10h30 -11h – Exemplos de uso: métodos e ferramentas do CGEE

11h – 12h – Análises de dados e RH para CTI

12h – 14h – Almoço

14h -14h30 – Análise de rede - principais aspectos teóricos e conceituais

14:30h – 15h – Características e funcionalidades básicas das ferramentas Gephi, insightNet e insightNet Browser

15h – 17h -- Exercício prático de uso do insightNet e insightNet

Browser

Roteiro 1 – currículos Lattes

Roteiro 2 – resumos WoS e Scopus

19/12

9h – 11h – Inteligência Tecnológica e insightData

11h – 12h – Demonstração do uso da insightData

12h – 14h – Almoço

14h -- 14h30 – Características e funcionalidades do VosViewer

14:30 -- 15h – Separação em grupos para exercício de planejamento de análise de redes

15h-17h – Exercício prático de análise exploratória de dados em redes

17h – 17h30 – Avaliação do treinamento e de percepção de potencial de uso das ferramentas no trabalho no CNPq

A seguir, são detalhados os principais conceitos, métodos e ferramentas exibidos.

3.2 Contexto metodológico da capacitação

Em articulação com os métodos qualitativos tradicionais, vários dos estudos realizados no CGEE são baseadas no que é conhecido como Ciência de Dados, que procura identificar variáveis e métricas que PODEM ser melhores descritoras ou preditoras de padrões existentes em dados. A aplicação da ciência de dados requer várias metodologias de áreas diferentes, desde o Processamento de Linguagem Natural e aprendizagem de máquina, até a análise de redes complexas e estatística, dentre outras. A forma como se aborda o problema de responder às perguntas norteadoras por meio das ferramentas e bases de dados disponíveis é crucial para que o melhor resultado final seja obtido.

Neste sentido, o objetivo do treinamento realizado foi fazer uma primeira exposição para analistas do CNPq dos métodos e ferramentas utilizadas pelo CGEE de modo a nivelar o conhecimento básico para a fase de planejamento a ser realizada subsequentemente. Os conceitos e descrições de métodos e ferramentas mais importantes da capacitação realizada são os seguintes:

Processo de análise de dados em CTI

Ao se estudar inteligência tecnológica, faz-se necessário o uso de alguma metodologia para que seja possível a identificação de uma pergunta norteadora e dos consequentes processos de reconhecimento de padrões que ajudem na elaboração da resposta ao problema elicitado. No CGEE, esse processo é dividido em 3 fases principais: o planejamento inicial (onde há várias discussões sobre o ponto de partida e sobre o que se deseja alcançar com os resultados); a preparação dos dados (que inclui sua aquisição, extração, tratamento e carga em banco de dados); e, por fim, a execução e comunicação dos padrões identificados. Ao fim deste processo, deseja se obter algum conhecimento acerca do tópico de interesse para que os atores envolvidos possam tomar decisões com base em evidências.

Inteligência Tecnológica e insightData

O foco das análises de dados realizadas pelo CGEE é a geração de informação útil para responder às perguntas norteadoras com o objetivo final de, juntamente com o cliente, consolidar o conhecimento acerca dos assuntos propostos. Essas análises podem ser inseridas no contexto de inteligência tecnológica, que consiste no processo sistemático de coleta, tratamento, análise e disseminação de informações sobre ciência, tecnologia e inovação visando subsidiar o processo decisório, o planejamento e a consecução de objetivos organizacionais. A inteligência tecnológica é construída em um ciclo de 5 etapas: planejamento e direção; coleta de dados; análise dos dados; disseminação do que foi aprendido; e avaliação das decisões tomadas.

Para auxiliar neste ciclo, no que diz respeito a dados de textos não

estruturados, o CGEE desenvolveu uma software denominada insightData, que permite explorar diversas bases de dados textuais disponíveis ao Centro e transformar tais dados para formatos que podem ser adicionalmente explorados em outras ferramentas de análise. Por meio do estudo dos conteúdos de documentos (tais como notícias, artigos e patentes), o analista pode recuperar informações, identificar padrões e monitorar tendências. Para tanto, a *insightData* tem funcionalidades para: organização visual termos de busca de tópicos em formato de árvore (taxonomia), contagem de incidência de termos no tempo, dentro do conjunto selecionado de textos, nuvens de palavras indexadas por frequência e relevância, extração de entidades nomeadas de dentro dos textos selecionados, além de cálculo e exportação de redes de similaridade entre documentos.

Análise de Redes

Um dos principais métodos utilizadas nas ferramentas do CGEE é a análise de redes, que busca modelar fenômenos relacionais, em especial sociais, de forma matemática a fim de extrair informações e facilitar tomadas de decisão acerca de sistemas complexos.

O objeto matemático que o fundamenta uma rede complexa é o grafo, que é composto de nós ligados por arestas. Para exemplificar, pode-se considerar uma rede de amigos de uma mídia social como o Facebook ou o Linkedin, onde cada nó é uma pessoa e as arestas são as relações de amizade entre essas pessoas. Esse tipo de representação não se limita a interações sociais, tendo sido empregado em contextos diversos e inusitadosem que as relações são tão ou mais importantes do que os atores, incluindo comunidades de animais, malhas de voos, reações bioquímicas e coautorias de artigos acadêmicos, como mostrado no capítulo anterior.

Apesar de simples, os grafos são instrumentos de análise muito úteis quando são levados em consideração sistema reais com muitos atores interconectados em estruturas complexas. Qualquer sistema real que pode ser modelado por um grafo é chamado de rede. Uma rede é dita complexa quando é possível identificar quatro características: componente conectada

gigante, distribuição de grau do tipo lei de potência, existência de comunidades e o efeito do mundo pequeno. Essas propriedades surgem espontaneamente no mundo real, mesmo sem a consciência dos atores neste processo. Um dos propósitos do treinamento foi expor aos analistas do CNPq casos de uso desse tipo de análise.

Ferramentas Gephi, insightNet e insightNet Browser

Para a realização de análise de redes, é quase sempre necessário utilizar computadores para o processamento dos grandes volumes de dados disponíveis. Para a computação desses dados de redes vários pacotes de software, tanto abertos como proprietários, têm sido desenvolvidos. No CGEE, optou-se como plataforma preferencial para análises de redes o software de código aberto Gephi. Devido à recorrência da necessidade de processar dados bibliométricos e da plataforma Lattes, da qual o Centro tem um espelho, foi desenvolvido internamente um *plugin* do Gephi, chamado de insightNet capaz de extrair e tratar esses dados. Além disso, para visualizações mais sofisticadas da informação processada pelo Gephi e pelo insightNet, foi desenvolvida a insightNet Browser, que funciona com base no navegador Firefox.

O Gephi é uma ferramenta de código aberto escrita em Java desenvolvida de maneira independente por diversos voluntários ao redor do mundo. O Gephi tem a vantagem de lidar bem com conjuntos de dados muito grandes, além de ser bastante amigável para usuários leigos, quando comparado com ferramentas similares.

Na versão atual, o insightNet. tem a funcionalidade de baixar dados da versão Lattes alojada no Portal Inovação, gerenciado pelo CGEE, carregar dados nos formatos do Scopus e do Web of Science e realizar diversas análises a partir de cálculos de métricas de redes próprias ou acessando as do próprio Gephi, Além disso, o insightNet pode coletar palavras-chave dos metadados de currículos e artigos, bem como produzir nuvens de termos com as palavras-chave coletadas, para mapeamentos temáticos dos conjuntos de dados selecionados.

A fim de facilitar a visualização dos dados, o CGEE também desenvolveu o insightNet Browser, onde o usuário pode, após realizar os cálculos necessários no Gephi e/ou no insightNet, visualizar os resultados em no Firefox, exibir nuvens de termos de vários tipos, produzir histogramas e gráficos de espalhamento dos metadados dos conjuntos de dados em análise, além de ter várias versões de filtros escolhidos visual e textualmente, muito úteis para análises exploratórias.

Recursos Humanos para Ciência, Tecnologia e Inovação (RH para CTI)

Para complementar o evento de capacitação, foram também apresentados os principais resultados das análises de dados de uma das atividades contínuas do CGEE, o Projeto “RH para CTI”, no qual são realizadas extensas análises de dados de egressos de programas de fomento à formação de pesquisadores em nível de graduação e pós-graduação, inclusive com bolsas do CNPq.

3.3 Benefícios esperados da capacitação

Como benefícios imediatos da capacitação realizada, espera-se:

1. Equipes preparadas para contribuir, em colaboração com a equipe do CGEE, na elaboração dos processos de análises de dados definidos para os programas selecionados;
2. Equipe capacitada de acordo com os processos definidos no uso de ferramentas e bases de dados disponíveis e de interesse do CNPq;
3. Desenvolver, em conjunto com analistas do CNPq, capacidade de contribuir na produção de protótipos de análises exploratórias para cada um dos programas selecionados;

Na última etapa da capacitação, foi feita uma discussão sobre os conteúdos apresentados na qual a própria capacitação foi avaliada bem como foram discutidas perspectivas de uso dos conteúdos apredidos.

Anexo A Slides selecionados das apresentações da capacitação

cgee | Centro de Gestão e Estudos Estratégicos
Ciência, Tecnologia e Inovação

Contexto: análise de dados para decisões baseadas em evidências

A ciência de dados procura identificar variáveis e métricas que PODEM ser melhores descritoras ou preditoras de padrões existentes em dados

Analisar dados demanda estratégia

Agrupar, inspecionar tipos de dados, limpá-los, determinar como os dados se associam, comparar, contrastar, encontrar similaridades e diferenças e, finalmente, encontrar sequências e padrões...

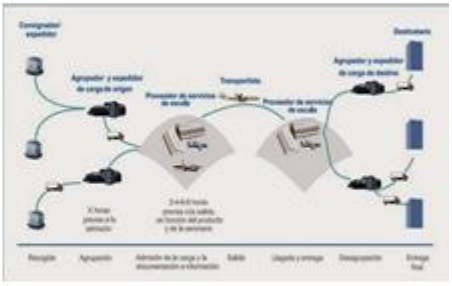
- É um processo extremamente trabalhoso e recorrente
- Tentativa, erro, frustração...
- Gratificante no final
- Mais fácil se houver um plano

A estratégia depende do contexto do "negócio"

Estratégia, bases de dados, ferramentas, infraestrutura... → análise de dados é um projeto

Dados como carga

x



Cada tipo define abordagens diferentes

Mesmo definido o negócio, não existe uma ferramenta que resolva tudo

Volume

Velocidade

Variedade

Tudo começa com uma boa definição de objetivos de “negócio” e dos atributos/variáveis antes de escolher bases e ferramentas de análise

Papéis

Especialistas de "negócio" – "Brainstorm" das variáveis que eles(as) acham que podem ser as melhores preditoras ou descritoras dos padrões a serem buscados

Cientistas de dados – Quantificam QUAIS das métricas e variáveis propostas SÃO as melhores preditoras ou descritoras dos padrões

Objetivos

- 1º Treinamento: Expor preliminarmente métodos e ferramentas utilizados pelo CGEE em bases de dados de CTI
- 2018: Juntar os dois times para desenhar e validar processos específicos de análise de dados que ajudarão a realizar o trabalho de fundamentar com evidências os cenários e avaliações pretendidas
- Realizar a transferência de conhecimento com base em métodos e ferramentas desenvolvidos pelo CGEE sobre os processos validados junto com o CNPq na construção de metodologias e processos de análise de dados aplicados à elaboração de cenários prospectivos de desenvolvimento institucional e avaliação dos programas INCT, Sisbiota Brasil e Ciências Ambientais

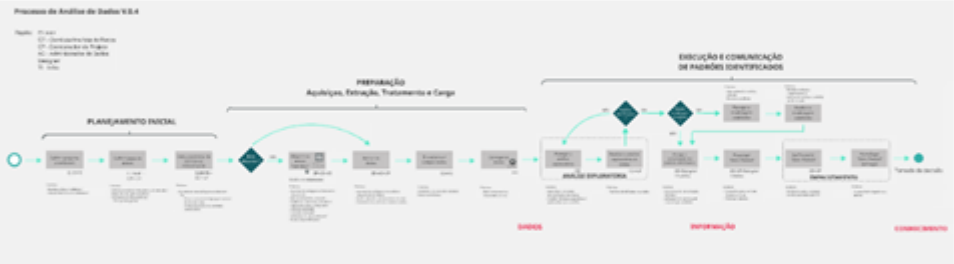
cgee | Centro de Gestão e Estudos Estratégicos
Ciência, Tecnologia e Inovação

CNPq- CGEE
Treinamento Básico - Inteligência Tecnológica
O processo de análise de dados em CTI

cgee

Processos de análise de dados

Processo básico de análise de dados



- Processos específicos têm que ser executados para cada pergunta
- O ETL normalmente consome 80-90% do tempo
- As ferramentas do CGEE aceleram MUITO o ETL e simplificam a análise exploratória de importantes fontes de dados de CTI

Natureza das fontes de dados

Plataforma Lattes

Scopus

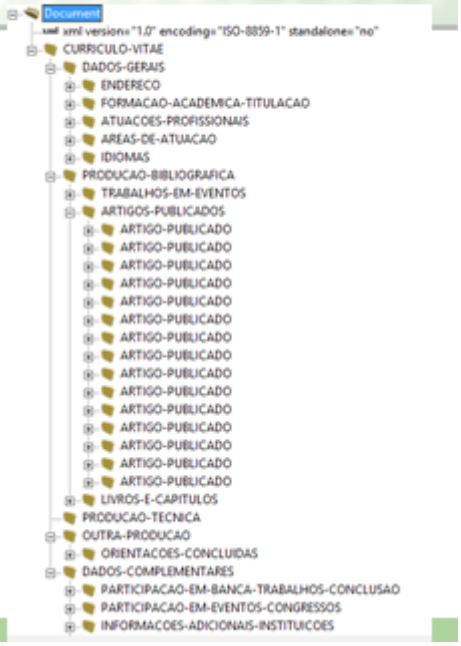
WEB OF SCIENCE™

uspto

Fontes CNPq

Plataforma Lattes

- 5 milhões de currículos
- Atualização diária
- Preenchido pelos usuários
- Acesso e extração via CGEE
- Identificação única por usuário
- DOI de publicações
- Formato XML
- Não tem os conteúdos das produções

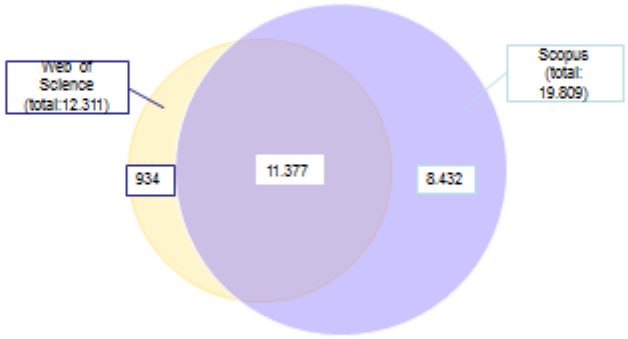


Bancos de dados bibliográficos

- Disponibilizam metadados de publicações e suas citações:
  - -autores
  - -títulos
  - -afiliações
  - -resumos
  - -ano de publicação
  - -DOI
  - -citações
  - -palavras-chave
  - -outros
- Requisito: acesso e extração via Portal Periódicos, da CAPES (CAFE/RNP)
- Usaremos aqui o Scopus, da Elsevier e o Web of Science, da Clarivate (Ex Thomson Reuters)

Slide 7

Scopus e Web of Science – Periódicos indexados (2014)



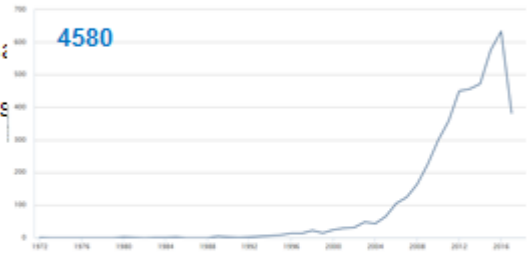
Slide 8

Scopus e Web of Science – Resultados práticos

Busca: "innovation" com afiliação no Brasil

Scopus:

- Mais publicações
- Ferramenta de busca mais versá
- 2000 metadados/download
- Viés para publicações europeias
- Dados a partir de 1972



Web of Science:

- Publicações de maior "qualidade
- Melhor para análise de citações
- 500 metadados/download
- Viés para publicações dos EUA
- Dados a partir de 1945 (ou antes

Slide 9

Alternativas metodológicas

Slide 11

Métodos e ferramentas

- Análise de redes
- Processamento de linguagem natural para mineração de textos
- Análise bibliométrica e mapas bibliográficos interativos
- Mapas de temas
- Exemplos na próxima apresentação

Insight
tools family

Centro de Gestão e Estudos Estratégicos
Ciência, Tecnologia e Inovação

CNPq- CGEE
Treinamento Básico - Inteligência Tecnológica
Análise de redes – principais aspectos teóricos e conceituais

Euler e a origem da teoria dos grafos

As 7 pontes de Königsberg (1736)

Mapa de Königsberg em 1651

Aqui nasceram:
Kant
Goldbach
Kirchhoff
Lipschitz
Clebsch
Hilbert
Sommerfeld
Wallach (N. Bioq.)
Lipman (N. Quim)
Arendt
Muitos outros...

Teoria dos grafos

- Pode ser usada em qualquer problema que envolva relações entre elementos
- Os elementos são representados por pontos (nós, ou vértices) e as relações são representadas por linhas entre dois pontos (arestas)
- Arestas podem ou não ser direcionadas, de modo que os grafos resultantes são chamados de não-direcionados ou direcionados (figs (a) e (b), respectivamente
- Arestas podem ter pesos (fig. (c))

Slide 4

O que são as relações?

Se os nós:

- Estão relacionadas por ligações simbólicas ou físicas (Exs. Facebook, Internet)
- Realizam os mesmos tipos de ações (Ex. cocitações)
- Dividem antecedentes (Ex. acoplamento bibliográfico)
- Possuem características comuns (Ex. Ingredientes em receitas)
- Ligam-se a entidades comuns (Ex. atores que contracenam)
- Realizam transações entre si (Ex. apertos de mão numa festa)
- São referidas como um par (Ex. relacionamentos amorosos)
- Referem-se a significados comuns (Ex. Ontologias, mapas mentais)
- Combinações das alternativas acima

Slide 6

De grafos a redes sociais

- Durkheim, Tönnies, Simmel teorizaram sobre descrições das relações sociais como redes
- Jacob Moreno (painel)

  - criador do psicodrama, pioneiro da terapia de grupo e da análise de redes sociais

  - usou grafos para representar relações entre pessoas (sociogramas) em "Who Shall Survive?" (1932)

Slide 8

Nós mais "importantes" da rede social?

- O nó com mais conexões – centralidade de grau
- O nó conectado aos nós mais conectados – centralidade de autovetor
- O nó que serve de "ponte", o corretor de conexões da rede – betweenness

Slide 9

Sistemas Complexos

Complex

[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]
–adjective
1.
composed of many interconnected parts; compound; composite: a complex highway system.
2.
characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.
3.
so complicated or intricate as to be hard to understand or deal with: a complex problem.

Fonte: Dictionary.com

Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

Fonte: John L. Casti, Encyclopædia Britannica

Complexity

cgee

Hipótese de trabalho:

Inerente a cada sistema complexo existe uma **rede** que descreve as relações entre seus componentes.

Redes complexas “naturais”

- Componente conectada gigante
- Distribuição de grau do tipo lei de potência
- Existência de comunidades
- Mundo pequeno

Condiciones sine quibus non

1. Que perguntas você espera responder?
2. O que são seus nós?
3. O que são suas arestas?
4. Como você extrairá os dados dos nós e arestas?
5. Qual o tamanho esperado da rede (número de nós e de arestas)?
6. Que impacto ou novidade você espera que o estudo dessa rede traga para o conhecimento existente sobre o(s) tema(s) escolhido(s)?

❑ As ferramentas desenvolvidas predefinem 2, 3 e 4 e exibem 5 de modo interativo e navegável (até o limite da capacidade do seu computador)

cgee | Centro de Gestão e Estudos Estratégicos
Ciência, Tecnologia e Inovação

CNPq- CGEE
Treinamento Básico - Inteligência Tecnológica

Gephi, insightNet, insightNet Browser

Gephi, iN, iN Browser

- Por que o Gephi?
- Gephi x Cytoscape, NodeXL, Pajek, Tulip ...
- Por que Gephi 0.8.2? → insightNet
- Integração com insightNet Browser
- Há muito a explorar
- Estamos perto de evoluir para o Gephi 0.9

Gephi makes graphs handy

Cytoscape

NODEXL

Tulip
Better Visualization Through Research

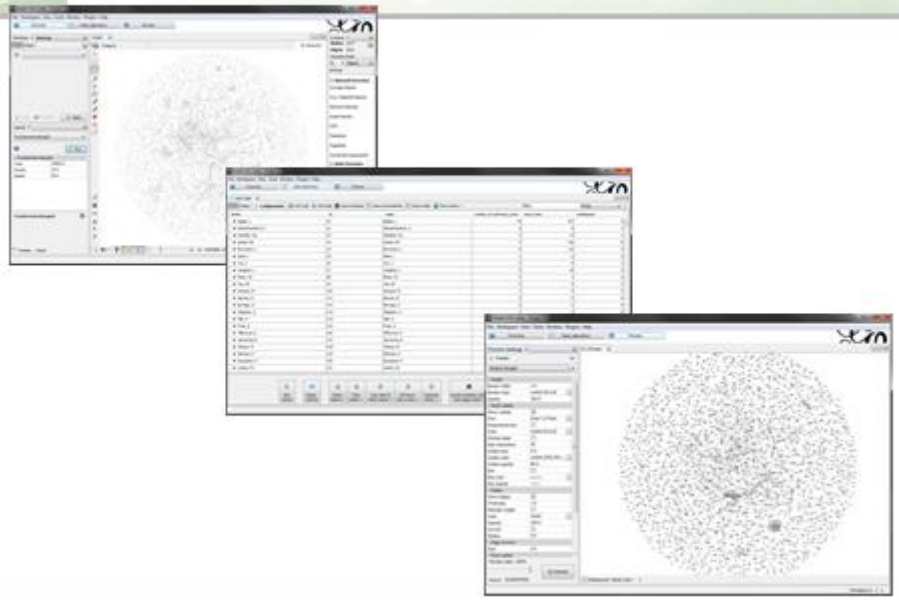
Slide 2

Gephi

- Ferramenta de código aberto concebida para exploração interativa e visualização de redes em tempo real
- Usa a placa gráfica do computador, liberando a CPU para processamento
- Escalável (até 30.000 nós de grau não nulo)
- Arquitetura multi-tarefa que aproveita processadores *multi-core*
- Pode carregar a maioria dos formatos de dados de redes (inclusive planilhas, no laboratório de dados)
- Voluntários oferecem uma extensa lista de plugins que extendem as funcionalidades da ferramenta

Slide 3

Interfaces de usuário



Slide 4

insightNet

- Criado para importar dados extraídos da plataforma Lattes e das bases bibliográficas Scopus e Web of Science
- O nós são currículos (Lattes) e artigos (Scopus e WoS)
- As arestas são coautorias ou similaridade semântica para dados do Lattes e apenas similaridade semântica para metadados do Scopus/WoS
- A ferramenta permite misturar dados do Scopus e do Web of Science, mas não é possível misturar dados do Lattes, no momento
- Foco na ANÁLISE EXPLORATÓRIA dos dados

Slide 6

Gerenciamento de dados no iN

- Os dados importados pelo iN ficam em um banco de dados SEPARADO do banco dados do Gephi
- Basicamente, dados importados dos currículos/artigos ficam exclusivamente no banco do iN enquanto dados das visualizações ficam exclusivamente no banco do Gephi
- O laboratório de dados (LD) integra os metadados dos nós e arestas dos dois bancos, mas com algumas limitações
- Dados das visualizações + LD podem ser exportados para o iN Browser no formato GEXF (de *Graph Exchange XML Format*)

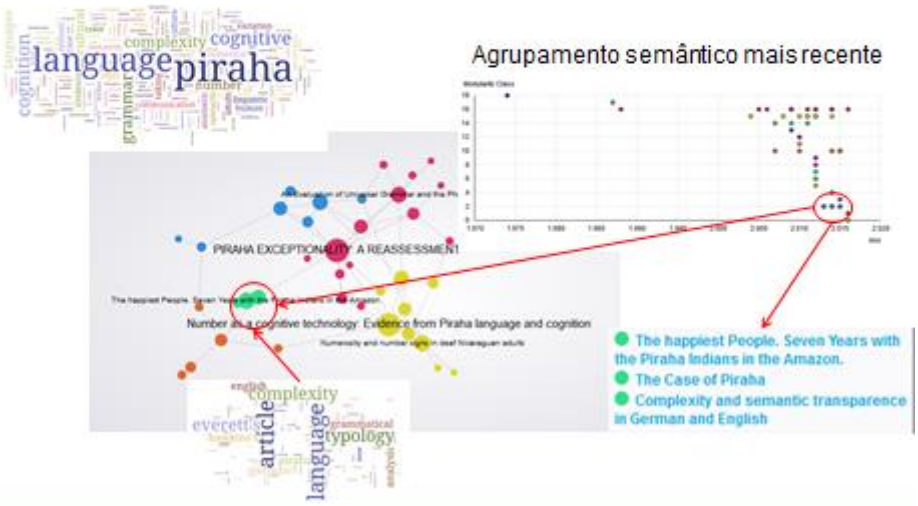
Slide 8

insightNet Browser

- Aplicação web para visualização e navegação em tempo real em redes exportadas no formato GEXF
- Portátil: funciona direto no navegador Firefox, sem necessidade de instalação no sistema operacional
- Permite a exploração das redes de forma interativa bem mais simples do que no Gephi/iN
- Expande as possibilidades de visualização, exploração e filtragem de metadados do Gephi/iN (nuvens de termos e gráficos de espalhamento)
- Exporta as visualizações e dados filtrados em vários formatos
- Foco na COMUNICAÇÂO dos dados: **só é exibido o que tiver sido preparado no Gephi e/ou no iN**

Slide 10

Artigos sobre a tribo Pirahã no WoS



Slide 12

Roteiro 1: Importação de rede de currículos

1. Escolher tema
2. Explorar opções da janela de importação a partir da base do CGEE, incluindo opções para rastreamento dos dados importados
3. Importar dados
4. Explorar janela de processamento das arestas
5. Processar arestas
6. Calcular centralidades e modularidade
7. Explorar partições
8. Explorar buscas por palavras-chave
9. Exportar tabela de palavras-chave
10. Exportar csv no laboratório de dados
11. Exportar gexf para o iN Browser
12. Abrir rede gexf gerado
13. Explorar nuvem de termos
14. Explorar gráficos de dispersão (*scatterplot*) e histogramas
15. Explorar currículos com o iNBrowser → abrir alguns currículos!

Roteiro 2: Importação de rede de artigos

1. Escolher tema
2. Entrar no Portal Periódicos
3. Entrar na página da base bibliográfica
4. Buscar artigos
5. Exportar artigos no formato BibTex
6. Explorar a janela de importação de BibTex do iN
7. Importar artigos
8. Repetir passos do Roteiro 2
9. Abrir algumas páginas web de artigos a partir do iN Browser

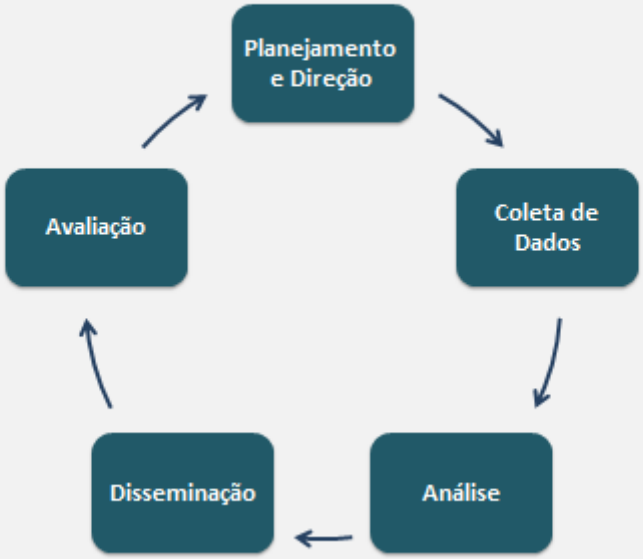
Inteligência Tecnológica e insightData

Eduardo Moresi
emoresi@cgee.org.br

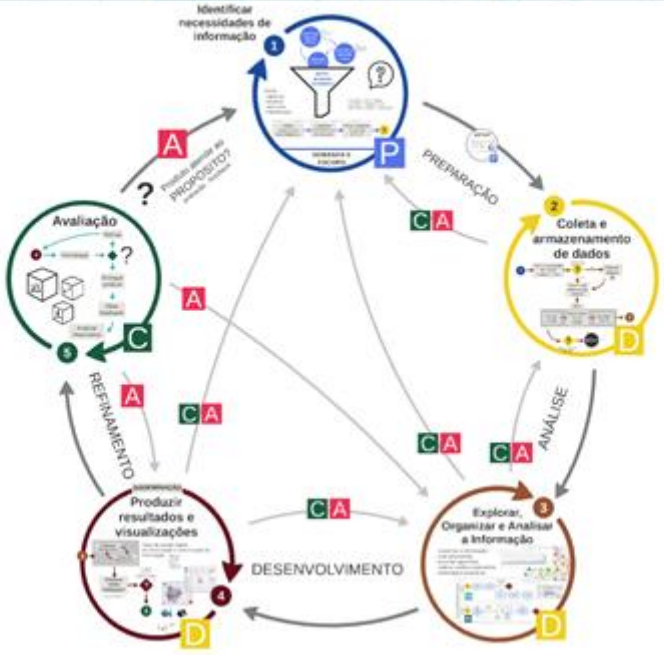
Rio de Janeiro, 18 de setembro de 2017

cgee

Inteligência Tecnológica



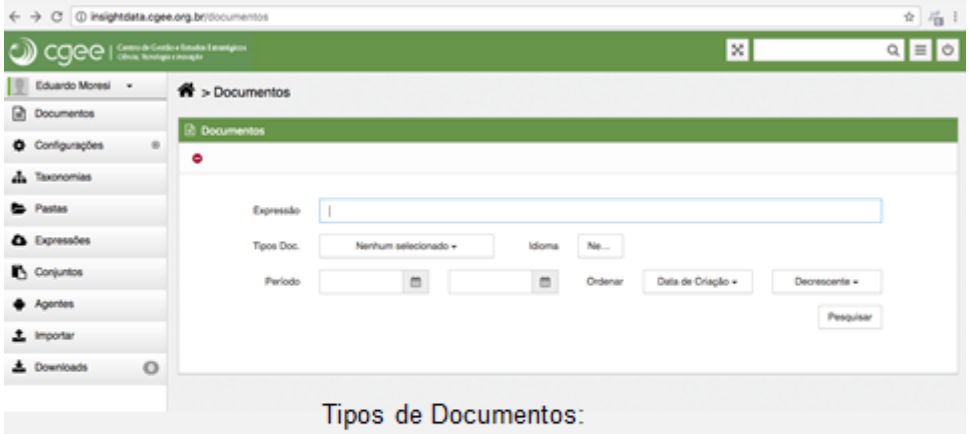
Ciclo de Inteligência



Características da insightData

- Memória organizacional: indexação, armazenamento e recuperação de grandes volumes de informações textuais;
- Apoio à recuperação de informações por tipo de fonte: artigos científicos, conteúdo noticioso, patentes e publicações;
- Apoia a identificação de padrões e interseções na recuperação de informações de diferentes áreas do conhecimento;
- Possibilita o monitoramento da evolução da frequência de termos para a análise de tendências e a identificação de sinais fracos.

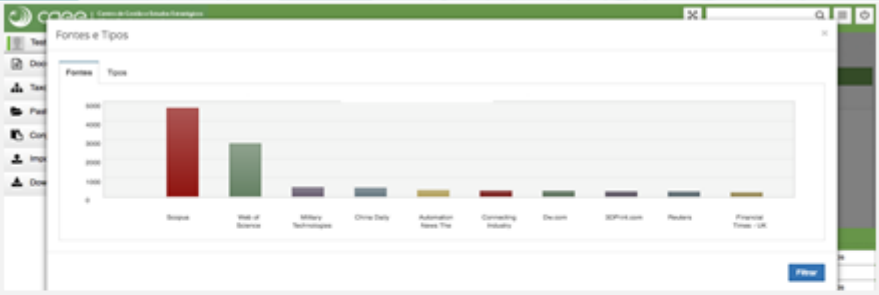
Funcionalidades da insightData



Tipos de Documentos:

- Notícias
- Acervo Scopus
- Acervo Web of Science
- Patentes USPTO
- Documentos CNPq

Curva de Fontes e Tipos




Entidades Nomeadas
> Entidades Nomeadas
Pesquisar

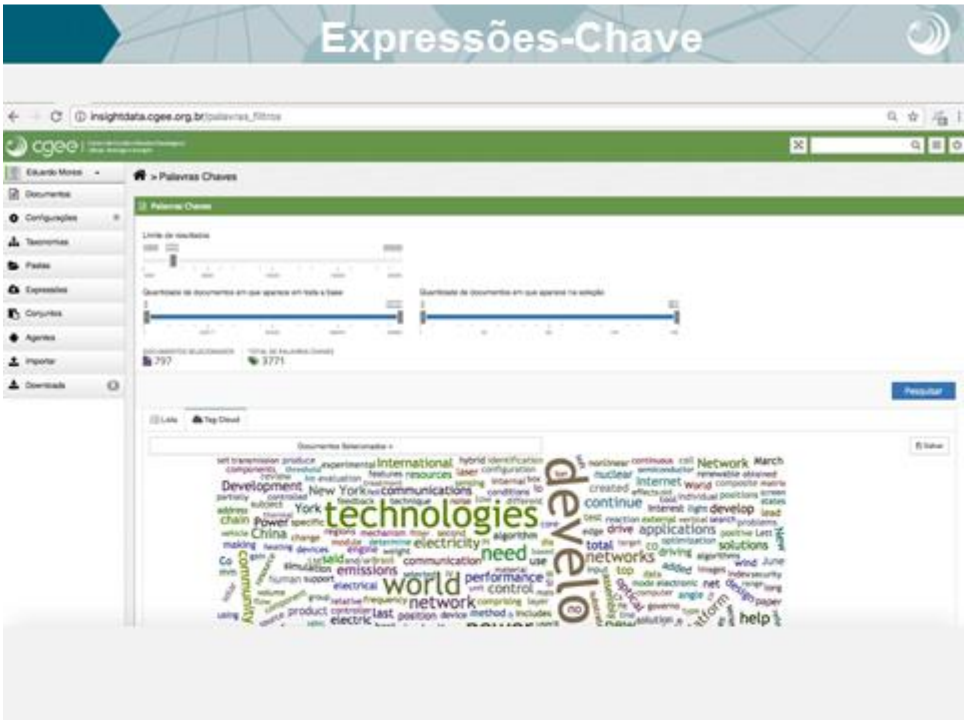
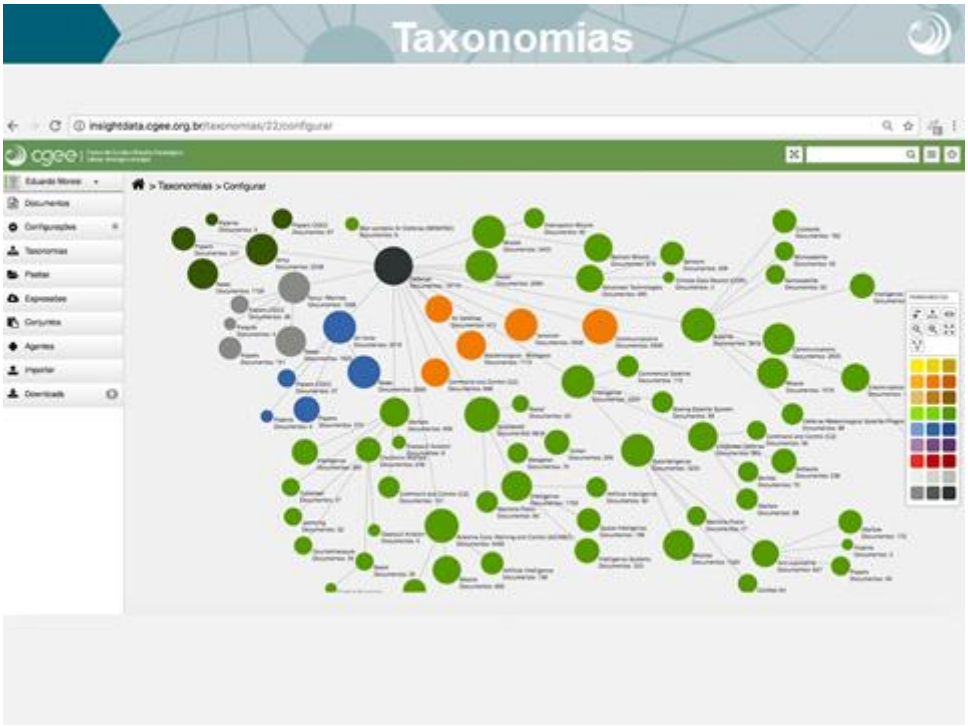

Expressões-Chave
> Palavras Chaves
Pesquisar


Gráfico de Termos
insightdata.cgee.org.br/documentos
Ocorrências em [todos os tipos de documentos]
advanced manufacturing
industry 4.0
low-carbon economy
production facility
Adicionar em Pasta
Crescimento de Termos
Curva de fortes/fipos
Entidades Nomeadas
Expressões Chaves
Gráfico de Termos
Exportar Documentos (XML)


Taxonomias
insightdata.cgee.org.br/taxonomias/22/configurar
> Taxonomias > Configurar
Documentos
Configurações
Taxonomias
Pastas
Expressões
Conjuntos
Agentes
Importar
Downloads