DOCUMENTOS

167

ISSN 1677-9274
Outubro2020

Representação do conhecimento sobre pragas e doenças do café: esquemas conceituais e recursos terminológicos

Embrapa
Informática Agropecuária

ISSN 1677-9274
Agosto/2020

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

DOCUMENTOS 167

Representação do conhecimento sobre as pragas do café: esquemas conceituais e recursos terminológicos

Ivo Pierozzi Júnior
Leandro Henrique Mendonça de Oliveira

Autores

Embrapa Informática Agropecuária
Campinas, SP
2020

Exemplares desta publicação podem ser adquiridos na:

Embrapa Informática Agropecuária

Av. Dr. André Tosello, 209 - Cidade Universitária
Campinas, SP, Brasil
CEP. 13083-886
Fone: (19) 3211-5700
www.embrapa.br
www.embrapa.br/fale-conosco/sac

Revisão de texto
Adriana Farah Gonzalez

Normalização bibliográfica
Carla Cristiane Osawa

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Marília Bastos sob supervisão de Magda Cruciol

Foto da capa
Ivo Pierozzi Júnior

1ª edição
Versão digital (2020)

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Informática Agropecuária

Pierozzi Júnior, Ivo,

Representação do conhecimento sobre pragas e doenças do café : esquemas conceituais e recursos terminológicos / Ivo Pierozzi Júnior, Leandro Henrique Mendonça de Oliveira. - Campinas : Embrapa Informática Agropecuária, 2020.

PDF (32 p.) : il. color. - (Documentos / Embrapa Informática Agropecuária, ISSN 1677-9274 ; 167).

1. Pragas. 2.Doenças do cafeeiro. 3. Terminologia. 4. Esquema conceitual. 5. Glossário.I. Oliveira, Leandro Henrique Mendonça de. II. Título. III. Embrapa Informática Agropecuária. IV. Série.

CDD (21. ed.) 410.855

Carla Cristiane Osawa (CRB-8/10421)

Autores

Ivo Pierozzi Júnior

Biólogo, doutor em Ecologia, pesquisador da Embrapa Informática Agropecuária, Campinas, SP

Leandro Henrique Mendonça de Oliveira

Cientista da computação, doutor em Ciências da Computação e Matemática Computacional, analista da Secretaria de Pesquisa e Desenvolvimento, Embrapa Sede, Brasília, DF

Apresentação

Neste documento é apresentado o relato de um trabalho de desenvolvimento de alguns formatos de representação do conhecimento sobre pragas e doenças do cafeeiro, realizado no âmbito do projeto "Tecnologia da Informação para o manejo integrado de doenças e pragas do cafeeiro: modelagem, representação do conhecimento e ferramentas computacionais de diagnóstico e alerta".

O processo de representação do conhecimento é um exercício multidisciplinar, que costuma ser amplo, diverso e rico em oportunidades metodológicas e tecnológicas, não se comprometendo com um formato final definitivo. O processo deve ser entendido como um exercício constante de reflexão cognitiva sobre o domínio a ser modelado e pode basear-se em ferramentas de modelagem conceitual ou computacional que permitam e respeitem o dinamismo e a plasticidade inerentes ao conhecimento científico, que evolui e se modifica constantemente assim como às linguagens a ele associadas. Ou seja, resultados desse tipo de processo devem ser entendidos e trabalhados como referenciais do conhecimento explícito capturado de corpus textual e preparados para evoluírem.

Os resultados obtidos, neste exercício, agregam valor ao conjunto dos outros resultados produzidos no mesmo projeto, à medida que os esquemas conceituais e recursos terminológicos desenvolvidos, tal como o glossário terminológico de pragas e doenças do café, possam ser utilizados metodológica e computacionalmente, como suporte: a) ao desenvolvimento de soluções para navegabilidade e acessibilidade em banco de dados específicos e na web; b) à compatibilidade e desambiguação terminológicas e à interoperabilidade semântica entre sistemas de informação relativos a essa temática; c) à implantação de sistemas de recuperação da informação; d) à realização de estudos de usuários para adequação de ferramentas e instrumentos de apoio à recuperação de informação na Embrapa.

Trata-se de uma contribuição de cunho teórico e metodológico para apoiar o desenvolvimento de pesquisas e inovações em representação do conhecimento para o domínio de manejo integrado de doenças e pragas do cafeeiro.

Sílvia Maria Fonseca Silveira Massruhá
Chefe-geral
Embrapa Informática Agropecuária

Sumário

Introdução

A proposta de representação do conhecimento sobre pragas e doenças do cafeeiro aqui apresentada baseou-se na realização sequencial das seguintes etapas operacionais:

1) Construção de base de dados terminológicos a partir de corpus textual técnico-científico sobre pragas e doenças do cafeeiro e de extração semiautomática de termos presentes no corpus.

2) Concepção e construção de esquemas conceituais de representação do conhecimento sobre pragas e doenças do cafeeiro.

a) Esquema conceitual genérico.

b) Esquema conceitual resultante de mineração de dados em corpus textual.

c) Esquema conceitual resultante do alinhamento semântico e terminológico trabalhados nos dois esquemas mencionados acima (alíneas a e b).

3) Construção e implantação computacional de um modelo de glossário específico para a temática sobre pragas e doenças do cafeeiro.

O processo de representação do conhecimento é um exercício multidisciplinar e costuma ser amplo, diverso e rico em oportunidades metodológicas e tecnológicas, não se comprometendo com um formato final definitivo. O processo deve ser entendido como um exercício constante de reflexão cognitiva sobre o domínio a ser modelado e pode basear-se em ferramentas de modelagem conceitual ou computacional que permitam e respeitem o dinamismo e a plasticidade inerentes ao conhecimento científico que evolui e se modifica constantemente assim como às linguagens a ele associadas. Dessa forma, os formatos de representação escolhidos e desenvolvidos no âmbito deste trabalho (terminologias, esquemas conceituais e modelo de glossário) devem ser entendidos e trabalhados como referenciais do conhecimento explícito capturado de corpus textual e preparados para evoluírem.

O referencial teórico e metodológico das análises conceituais textuais, realizadas no contexto deste trabalho e que também embasam a construção de terminologias, encontra-se nas interfaces das Ciências da Cognição, da Informação e da Comunicação e da Linguística, em particular na Teoria do Conceito (Dahlberg, 1978) e na Teoria Comunicativa da Terminologia (Cabré, 1993, 1999, 2008). Os resultados das análises conceituais expressam-se em conjuntos de termos ou expressões que representam conceitos, os quais podem ser interrelacionados indefinidamente, obedecendo à normalização de um vocabulário controlado e agregando valor cognitivo para representação simplificada de conceitualizações complexas que, no caso particular desse trabalho, envolvem as interfaces dos domínios da Agronomia, Ecologia, Biologia, Zoologia, Fitotecnia, entre outras disciplinas. A extração e organização dos termos e expressões pode ser intelectual, semiautomatizada ou mesmo mista, agregando-se ambos os métodos, com suporte tecnológico de software de mineração de dados terminológicos, de edição de sistemas de organização do conhecimento e esquemas conceituais relacionais (Zeng, 2008; Souza et al., 2010; Pierozzi Junior et al., 2014; Netto et al., 2016).

Este documento está organizado da seguinte maneira: na presente seção, Introdução, apresentam-se os contextos nos quais o exercício de representação do conhecimento se configurou, com apresentação do referencial teórico que contextualiza as terminologias e os glossários como recursos de representação do conhecimento. Na Seção 2, apresentam-se os métodos, técnicas e tecnologias reunidos, organizados e utilizados na concepção e elaboração das terminologias e glossário. Na

Seção 3, apresentam-se os resultados obtidos, assim como suas formas e objetos de entrega. Na Seção 4, são feitas considerações finais sobre os resultados apresentados além de discutir-se perspectivas de evolução metodológica, tecnológica e processual para os mesmos.

1.1. Representação do conhecimento baseada em linguística de corpus textual: terminologias, esquemas conceituais e glossário

O conhecimento é uma experiência cognitiva humana e individual que, posteriormente, por meio de processos comunicacionais pode ser socializado, por exemplo, por meio de publicações textuais impressas ou digitais. A construção do conhecimento envolve a interação de vários elementos intelectuais, tais como inteligências e aprendizagem. A rigor não existe "transferência" de conhecimento, uma vez que sendo uma experiência individual ligada a uma enormidade de características e propriedades pessoais não pode ser simplesmente repassado na integridade de sua essência de um cérebro emissor para outro receptor. O conhecimento então é codificado por meio de signos e assim se transforma em um formato material, tangível, passível de comunicação interpessoal. O ser humano faz isso tão naturalmente que tal processo complexo se torna quase imperceptível. Um dos signos mais utilizados é a linguagem falada ou escrita, ou seja, a língua ou idioma por meio do qual o indivíduo detentor do conhecimento se comunica. Então, o conhecimento tácito se transforma em palavras, frases, textos e ganha possibilidades de interlocução e de novas significações.

O conhecimento acadêmico não se processa de forma diferente e sua "transferência", ou disseminação, obedece aos mesmos princípios genéricos descritos acima. A reunião e compilação de material textual que codifica por meio de linguagem escrita o conhecimento de determinado domínio ou especialidade configura-se um corpus textual e representa uma fonte imensamente rica de conhecimento a ser explorada e reusada.

Formalmente, é a Linguística de Corpus que organiza as bases conceituais e teóricas e o ferramental para o trabalho de exploração dos *corpus textuais:*

[...] *A Linguística de Corpus se ocupa da coleta e exploração dos corpora , ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador.* (Sardinha, 2000, p. 325)

A exploração dos corpora[1] científicos ou acadêmicos e a organização de dados textuais ali contidos revela o vocabulário ou o léxico da especialidade ou do domínio envolvido. Esse vocabulário, devidamente trabalhado pode constituir a terminologia específica daquele assunto e, como tal, converte-se em um recurso de representação do conhecimento ali explicitado. Dessa forma, as terminologias também podem ser consideradas como um recurso de representação do conhecimento e, do ponto de vista da semiótica baseada nos signos linguísticos, se coloca no início de um itinerário progressivo que reúne e organiza vários métodos e ferramentas, coletivamente chamados Sistemas de Organização do Conhecimento (SOC[2]) (Zeng, 2008).

[1] Corpora: plural de corpus.

[2] Tradução para o português do termo em inglês Knowledge Organization Systems (KOS), que são recursos originados das linguagens documentárias da Ciência da Informação e de ampla aplicação na Ciência da Computação.

1.2. Terminologias

A Terminologia pode ser considerada como a parte da Linguística que se dedica à linguagem das especialidades, ou seja, a linguagem praticada nos diversos segmentos do conhecimento humano (assim como de suas intermináveis subdivisões ou interseções), que se tornam específicos por focarem determinadas áreas como ciências, artes, técnicas ou profissões. Nesse aspecto reconhece-se que dentro de uma determinada área de conhecimento desenvolve-se uma linguagem cujo conjunto léxico torna-se voltado para a representação dos conceitos especificamente tratados naquele recorte de conhecimento.

O próprio termo "terminologia" é polissêmico, ou seja, tem mais de um significado. Segundo reiteram Almeida e Correia (2008) e Maculan (2015), baseadas em Cabré (1995), podem-se mencionar as seguintes acepções: a) o conjunto de postulados ou fundamentos teóricos necessários para dar suporte à análise de fenômenos linguísticos referentes à linguagem e comunicação especializada, assim constituindo-se a Terminologia como uma atividade, disciplina ou área de conhecimento e, então, grafada com a inicial maiúscula; e b) o conjunto vocabular, ou o conjunto dos termos, que se elege como componentes do léxico para um determinado domínio, nesse caso grafada com a inicial minúscula.

No seu papel de representação do conhecimento, reconhecem-se duas propriedades das terminologias: a) descrição e ordenamento do conhecimento, conferindo-lhes a capacidade de atuação no nível cognitivo, considerando que os termos denotam os conceitos[3] , conforme estabelece a Teoria do Conceito (Dahlberg, 1978); e b) compartilhamento e disseminação do conhecimento, conferindo-lhes a capacidade de atuação no nível comunicacional.

No contexto deste trabalho, a abordagem terminológica desenvolvida, além de ir ao encontro das duas propriedades acima mencionadas (facilitação cognitiva e comunicacional da conceitualização) foi dirigida também para a elaboração de um modelo de glossário, ação cuja execução é detalhada na seção 3.6.

1.3. Esquemas conceituais

No contexto do presente trabalho, tanto as terminologias como os esquemas conceituais são considerados como recursos de representação do conhecimento diretamente originados de processos cognitivos naturais humanos, especificamente a linguagem natural escrita. Por meio de ferramentas de Processamento de Linguagem Natural (PLN) e de ferramentas que visam a transcrição do conhecimento para linguagens artificiais ou "linguagens de máquina", tanto as terminologias como os esquemas conceituais são modelados objetivando agregar valor à gestão da informação e do conhecimento quando tais recursos são alinhados a sistemas de informação referentes ao domínio envolvido.

No caso específico do domínio de conhecimento sobre pragas e doenças do cafeeiro, o objetivo de desenvolvimento dos esquemas conceituais foi oferecer suporte operacional para melhoria de desempenho de banco de dados ou quaisquer outros tipos de sistemas de informação sobre essa temática que eventualmente possam vir a ser construídos ou mesmo aos sistemas de informação já existentes e que possam ter melhor desempenho em suas funções.

[3] O conceito é considerado a unidade do conhecimento (Dahlberg, 2014).

Nesse contexto, é oportuno apresentar como os esquemas conceituais podem ser construídos e utilizados. Para isso, o referencial teórico e metodológico apresentado por Olivé e Cabot (2007) foi adotado e é sucintamente apresentado a seguir. Segundo os autores, sistemas de informação (SI) são concebidos e executados para funcionarem como:

1) Memória: para manter a representação do estado de um domínio.

2) Informativo: para prover informação sobre o estado do domínio.

3) Ativo: para desempenhar ações que mudem o estado do domínio.

No processo de modelagem conceitual de um Sistema de Informação (SI), para que o mesmo seja capaz de desempenhar aquelas funções, o sistema requer conhecimento sobre o domínio e sobre as ações que ele tem que desempenhar. Genericamente, assume-se que um domínio consiste de um número de objetos, os quais são classificados em conceitos, e dos respectivos e consequentes relacionamentos entre eles.

A suposição de que um domínio consiste de objetos, relacionamentos e conceitos é uma maneira específica de visualizar o mundo. À primeira vista, parece uma suposição óbvia. A verdade da questão é, no entanto, bastante diferente. Existem outras opiniões possíveis que podem ser mais adequadas em outros campos. Para dar um exemplo simples e bem conhecido, em lógica proposicional assume-se que os domínios consistam de fatos, que podem ser verdadeiros ou falsos. O estudo da natureza e organização do mundo real é um ramo de filosofia chamado “ontologia”. Quando se assume que um domínio consiste de objetos (conceitos) e seus relacionamentos, surge um comprometimento com uma maneira específica de representação dos domínios. O termo usado na ontologia para denotar esse compromisso é “compromisso ontológico”. Na área de SI, esse compromisso de representação particular é chamado de “modelo conceitual”. O conjunto de conceitos utilizados em um domínio particular constitui a “conceitualização” do domínio. Enquanto na Filosofia entende-se ontologia como a “constituição do ser”, em Computação ontologia é a “especificação de uma conceitualização”, tornando-se uma visão concreta de um determinado domínio. Na área de SI, ontologias são chamadas de “esquemas conceituais” e as linguagens nas quais eles são escritos são chamadas de “linguagens de modelagem conceitual”. Dessa forma, as ontologias (especificações de um determinado domínio) podem ser convertidas em artefatos computacionais e utilizadas como suporte operacional em SI para organização e recuperação de informações, por exemplo.

1.4. Glossários

Assim como as terminologias, os glossários também podem ser considerados como SOC (Souza et al., 2010), na medida em que agregam valor às terminologias com a proposição de definições ou das acepções para cada conceito/termo e, uma vez mais, reforçando as propriedades cognitivas e comunicacionais da semiótica linguística.

Um glossário é uma lista alfabética de termos de um domínio de conhecimento específico, acompanhados de suas respectivas definições. Originalmente as anotações sobre o sentido das palavras (glosas) eram interlineares ou colocadas nas margens dos textos para esclarecer o sentido de palavras menos conhecidas; mais tarde, a proposta dos glossários evoluiu para a parte final do texto ou mesmo para um volume próprio agregado à obra à qual se referia, constituindo-se um suporte ao seu entendimento.

Glossários podem ser mono, bi ou multilíngues, sendo que estes últimos apresentam os termos em uma determinada língua acompanhados das indicações de seus respectivos equivalentes em outros idiomas.

O verbete, ou seja, o trecho textual que apresenta o termo e sua definição, em lexicografia, é o conjunto das acepções, exemplos e outras informações pertinentes, contido numa entrada de dicionário, enciclopédia ou glossário.

Assim compreendidos e constituídos, os glossários justificam-se como ferramentas de representação do conhecimento na medida em que sejam desenvolvidos em sintonia, sincronia e sinergia com o conteúdo informacional (conhecimento codificado em linguagem natural e/ou artificial) pois, assim, comprometem-se com a manutenção da consistência cognitiva e comunicativa dos domínios de conhecimento por eles representados.

No contexto do presente trabalho, o modelo de glossário proposto para pragas e doenças do cafeeiro não restringe o conteúdo para apenas a terminologia agronômica. Baseado nos esquemas conceituais construídos, inclui-se a possibilidade de elencar também as entidades nomeadas mineradas do *corpus* textual como, por exemplo, pessoas e instituições que, ao invés de uma "definição" podem ser apresentadas por meio de informações como nome completo e competências profissionais, no caso de pessoas (especialistas do domínio) ou pelo endereço, siglas ou acrônimos, no caso de instituições. Dessa forma, além da propriedade usualmente atribuída aos glossários (definições técnicas dos conceitos do domínio), o recurso aqui proposto pode funcionar igualmente como uma lista de autoridades ou um catálogo descritivo de entidades emergidas da ontologia.

2. Material e métodos

2.1. Construção do corpus textual e extração semiautomática dos candidatos a termos

O *corpus* textual construído para execução do presente trabalho reuniu os textos de 306 artigos apresentados no "Simpósio de Pesquisa dos Cafés do Brasil", no período de 2000 a 2013. A seleção dos textos foi feita por especialistas no domínio do conhecimento de Manejo Integrado de Pragas (MIP) e de pragas e doenças do cafeeiro, considerando-se o escopo e a diversidade temática dos documentos. Na sequência foi realizada uma extração automática de termos, cujo resultado foi submetido a nova avaliação pelos especialistas de domínio e, com base nos índices de frequência e coocorrência dos termos, o corpus foi então validado como representativo do domínio específico de "pragas e doenças do cafeeiro".

Essa conduta metodológica foi adotada para evitar problemas operacionais relativos ao processamento e compilação computacionais de corpora textuais muito volumosos. Como na primeira fase a abordagem é estatística, a análise de subconjuntos pode revelar-se suficiente para a identificação e seleção de candidatos a termos, já que os principais parâmetros de seleção são frequência e coocorrência. Então, a partir de um certo volume de texto processado, a quantidade e relevância dos termos extraídos pode ser suficiente para o exercício de representação do domínio. A decisão de corte atribuída à validação realizada por especialistas torna o processo semiautomatizado, agregando-se à análise estatística automática uma garantia intelectual e cognitiva (conhecimento tácito) à terminologia. Além disso, também existe a inerente garantia literária, uma vez que os termos são extraídos de artigos publicados (conhecimento explícito) e que passam por avaliação por pares no

processo normalmente empregado de seleção dos trabalhos a serem apresentados em eventos técnico-científicos.

Na sequência, o *corpus* foi submetido ao processamento incluindo: a) limpeza dos textos (eliminação de "sujeiras" geradas na conversão de arquivos do formato PDF para TXT, como número de páginas, cabeçalhos, títulos de seções do documento, tabelas, figuras, etc.); b) reunião e compilação dos textos individuais para um arquivo TXT único; c) extração automática dos candidatos a termos; e d) mais uma etapa de validação por especialistas.

2.2. Concepção e construção das bases de dados terminológicos

As bases de dados terminológicos produzidas no presente trabalho são de duas naturezas:

1) Módulo corpus textual: trata-se de um arquivo em formato TXT com um texto único, reunido, limpo e compilado de 306 documentos, considerados qualitativamente relevantes para a representação do domínio de conhecimento sobre pragas e doenças do cafeeiro. Esses documentos foram selecionados de um conjunto maior reunindo 2.587 resumos apresentados no "Simpósio de Pesquisa dos Cafés do Brasil", no período de 2000 até 2013 representando, assim, aproximadamente 12% do volume textual original de referência.

2) Módulo vocabulário: trata-se da lista de termos gerada pelo processo semiautomatizado descrito acima (seção 2.1.1.), contendo aqueles selecionados estatisticamente e validados por especialistas e a partir da qual foram realizadas as etapas seguintes do trabalho: concepção e construção dos esquemas conceituais de representação do conhecimento e do modelo de glossário.

2.3. Ferramentas para concepção e construção dos esquemas conceituais de representação do conhecimento

Como já foi apresentado e discutido anteriormente (seção 1.3.), no contexto deste trabalho, o conceito de "esquemas conceituais" é estendido para um conjunto de formatos diagramáticos utilizados em processos de representação do conhecimento e que englobam: mapas mentais ou conceituais, grafos e redes, entre outras denominações encontradas nos referenciais teóricos e na literatura.

A concepção e construção dos esquemas conceituais, depois de trabalhadas em nível de ideias, se beneficia grandemente de ferramentas ou softwares de edição gráfica, os quais serão detalhados na seção 2.2. Cabe, aqui, esclarecer que, em sua maioria, essas ferramentas consistem em recursos para desenhos e arquiteturas de nós (ou nodos) e seus respectivos inter-relacionamentos (arestas) no espaço. Nas ações de representação do conhecimento, os nós desses esquemas conceituais representam conceitos componentes do domínio a ser modelado e as arestas que os unem representam os vários tipos de relações que se estabelecem entre os nós, notadamente, as hierarquias, as equivalências e as relações associativas livres.

Os nós (conceitos) que formam o esquema conceitual podem ser selecionados por processos de mineração de textos e extração de candidatos a termos[4] . As relações, ou significados que ligam dois conceitos no contexto da representação do conhecimento são determinados por rótulos se-

[4] A partir da Teoria do Conceito (Dahlberg, 1978) toma-se o "termo" como a forma lexicográfica denotativa do conceito

mânticos, os quais podem ser livremente criados ou reusados de outros esquemas conceituais como tesauros, redes semânticas ou ontologias.

No processo de concepção e construção dos esquemas conceituais do presente trabalho foram reusados os recursos conceituais e terminológicos do tesauro brasileiro Thesagro[5] e do tesauro multilíngue Agrovoc[6] , alinhados ao referencial teórico sobre Sistemas de Organização do Conhecimento (Zeng, 2008; Souza et al., 2010).

2.4. Mapeamento semântico complementar

Visando enriquecer a terminologia sobre pragas e doenças do cafeeiro que, no contexto desse trabalho, a técnica de mapeamento semântico foi empregada visando dois objetivos:

1) Mapear na literatura internacional, em língua inglesa, termos referentes a pragas e doenças do cafeeiro no Brasil, obtendo-se assim garantia literária para:

a) Os termos equivalentes em língua inglesa para os termos em língua portuguesa;

b) A relevância dos termos extraídos do corpus textual em língua portuguesa e representativo do domínio do conhecimento sobre a temática em nível nacional.

2) Testar os algoritmos de extração automática de termos e de visualização do mapa semântico do software VOSViewer, usando-o para parametrizar as análises linguísticas executadas conforme foi descrito na seção 2.1.1.;

O itinerário metodológico para o mapeamento semântico é descrito em seguida, e consiste nas etapas de: 1) geração dos arquivos de dados a serem importados no software VOSViewer, que constituem um novo corpus textual a ser trabalhado; e 2) importação dos arquivos de dados para análise e visualização no software VOSViewer:

1. Geração dos dados:

a) Bases de dados bibliográficos consultadas: Scopus (SCO) e Web of Science (WoS);

b) Expressão de busca: “coffee pest” OR “coffee diseases” AND Brazil;

c) Salvamento dos arquivos de dados:

i) SCO: *.RIS;

ii) WoS: *.TXT;

2. Análise dos dados: parâmetros configurados no software VOSViewer[7] para a visualização dos mapas:

a) SCO:

i) *Type of analysis: co-occurrence; counting method: full counting; unit of analysis: all keywords;*

ii) *Choose threshold: mininum number of occurrence of a keyword: 1;*

[5] Disponível em: <http://enagro.agricultura.gov.br/glossario/thesagro-pagina>.

[6] Disponível em: <http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>.

[7] Os parâmetros são apresentados em inglês, conforme estão descritos no software, que não possui versão para a língua portuguesa, permitindo o leitor identificar as escolhas realizadas para execução do presente trabalho.

iii) *Choose number of keywords: number of keywords to be selected: all;*

b) WoS:

i) *Type of analysis: co-occurrence; counting method: full counting; unit of analysis: all keywords;*

ii) *Choose threshold: mininum number of occurrence of a keyword: 1*

iii) *Choose number of keywords: number of keywords to be selected: all;*

2.5. Concepção e construção do modelo de glossário

A concepção e construção do modelo de glossário para pragas e doenças do cafeeiro foi baseada na metodologia "OntoMethodus", apresentada e discutida por Di Felippo et al. (2008). O OntoMethodus é uma metodologia para construção de ontologias especialmente de fontes de dados não estruturados como textos e se alinha às etapas metodológicas do software e-Termos[8] , utilizado para a execução da gestão terminológica adotada neste trabalho (v. detalhes sobre o e-Termos na seção 2.2). A Figura 1 apresenta a correlação entre as etapas metodológicas propostas pelo OntoMethodus e pelo e-Termos e que deu suporte ao processo de representação do conhecimento sobre pragas e doenças do cafeeiro proposto no presente trabalho.

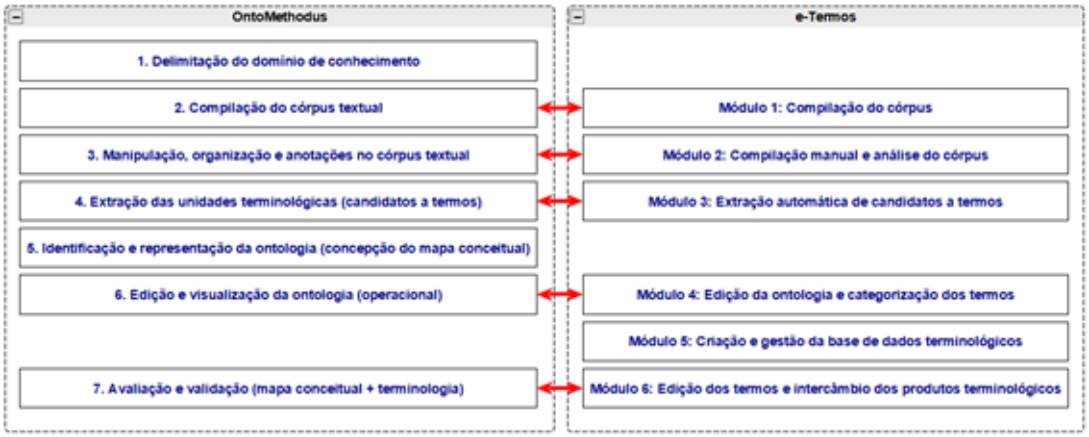

Figura 1. Correlação entre as etapas metodológicas do OntoMethodus (Di Felippo et al., 2008) e os módulos de trabalho do e-Termos8, usualmente empregadas em gestão terminológica.

O modelo da ficha terminológica para organização dos dados pertinentes de cada verbete do glossário é apresentado na Figura 2.

[8] Disponível em: <https://www.etermos.cnptia.embrapa.br/index.php>.

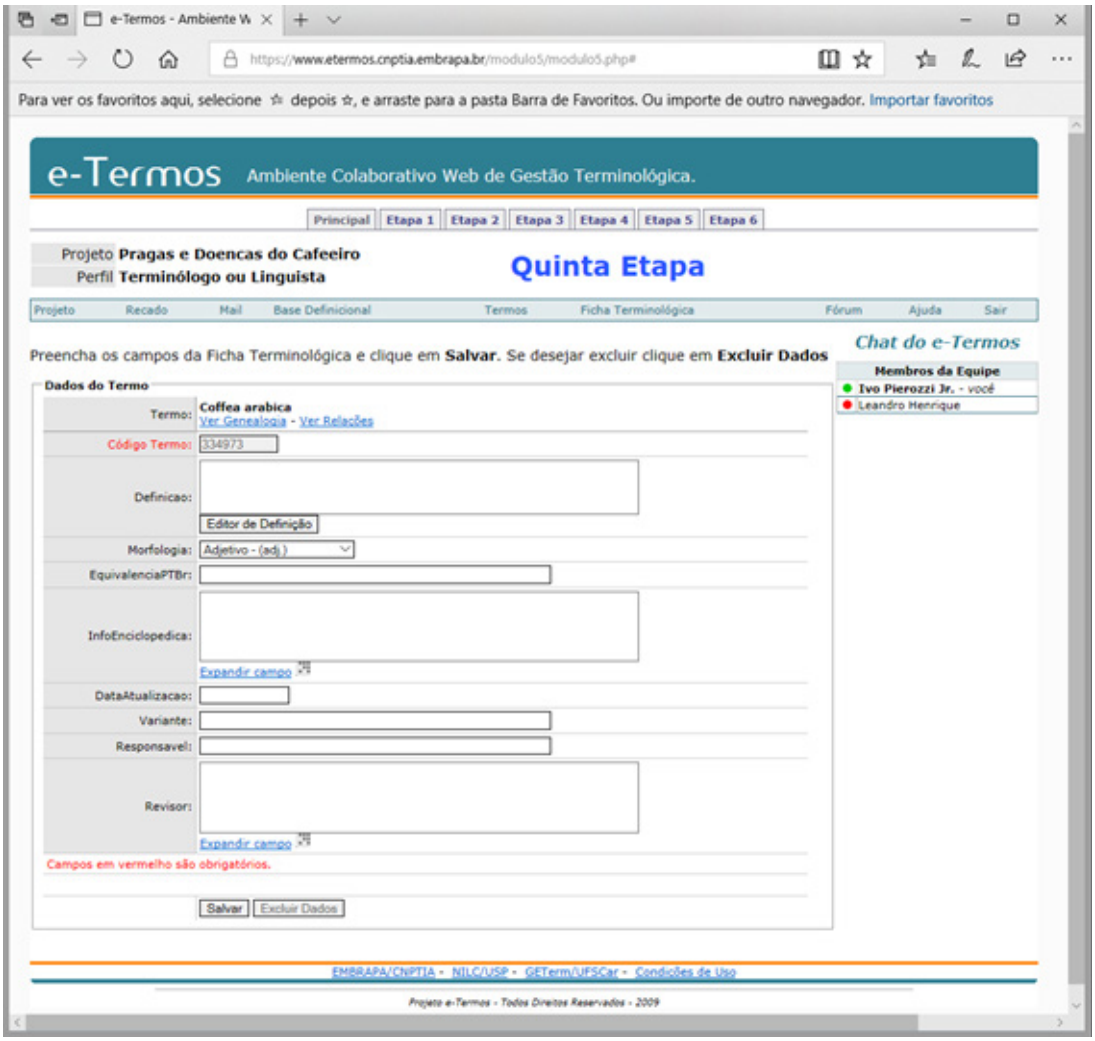

Figura 2. Modelo de ficha terminológica proposta para a descrição dos verbetes do glossário sobre pragas e doenças do cafeeiro. A ficha terminológica corresponde à Etapa 5 do software e-Termos[9]).

Nessa ficha terminológica foram previstos campos que comportam, para cada termo com interesse de ser transformado em verbete de glossário, os seguintes metadados:

CAMPO	CONTEÚDO
Termo	Formato lexográfico adotado terminológico, baseado na lista de candidatos a termos extraída do *corpus* textual
Código Termo	Referência numérica de ordenação do termo dentro do repositório de dados terminológicos
Definição	Enunciado que parafraseia a acepção de uma palavra ou locução pela indicação de suas características genéricas e específicas, de sua finalidade, pela sua inclusão num determinado campo do conhecimento etc.
Morfologia	Estudo da constituição das palavras e dos processos pelos quais elas são construídas a partir de suas partes componentes, os morfemas
InfoEnciclopedica	Trechos textuais com a menção dos termos que podem auxiliar na sua compreensão, indicar novas acepções e, assim, dar suporte ao enunciado definitório.
DataAtualização	Data na qual o terminológo ou revisor trabalhou no conteúdo da ficha terminológica
Variante	Sinonímias do termo
Responsável	Registro do nome do terminológico componente da equipe de gestão terminológica
Revisor	Registro do nome do revisor componente da equipe de gestão terminológica

[9] Disponível em: <https://www.etermos.cnptia.embrapa.br/index.php>.

2.6. Softwares

As seguintes etapas metodológicas foram realizadas por meio da utilização do software e-Termos[10] , que é um ambiente computacional colaborativo web de acesso livre e gratuito dedicado à gestão terminológica:

1) Concepção, construção, limpeza e compilação do corpus textual.

2) Extração semiautomática de candidatos a termos.

3) Armazenamento das bases de dados terminológicos.

Análises complementares do corpus foram realizadas por meio da plataforma Sketch Engine

Os esquemas conceituais foram construídos por meio da utilização do editor de diagramas yEd[11] e de uma versão especialmente adaptada para o presente trabalho do minerador de texto Sobek[13] .

O mapeamento semântico complementar foi executado por meio do software VOSViewer[14] .

Para maiores informações sobre os softwares utilizados recomenda-se o acesso e a navegação nos websites indicados nas respectivas notas de rodapé, onde detalhes técnicos, escopo e funcionalidades, assim como aplicabilidade de cada uma dessas ferramentas são apresentados, discutidos e exemplificados, no âmbito de sua utilização nas áreas de Processamento de Linguagem Natural, linguística de corpus, organização, representação, visualização e mapeamento de conhecimento.

3. Resultados

3.1. Corpus textual e bases de dados terminológicos sobre pragas e doenças do cafeeiro

A análise do corpus textual contou 1.132.082 palavras sendo 46.461 formas de palavras únicas. Entre essas a Tabela 1, de modo ilustrativo, apresenta as 50 palavras mais frequentes, respectivamente, palavras simples = unigramas e palavras compostas = bi, tri, tetra, etc.-gramas). Esse primeiro resultado de extração automática demonstra a necessidade de uma avaliação intelectual, pois nem todas as palavras extraídas serão consideradas nas etapas subsequentes, como é o caso de “were”, “was”, “of coffee”, entre outras.

Tabela 1. As 50 palavras mais frequentes extraídas automaticamente do corpus textual sobre pragas e doenças do cafeeiro.

Palavras simples			Palavras compostas	
No.	Termos	Frequência absoluta	Termos	Frequência absoluta
1	cafeeiro	5667	coffea arabica	619
2	coffea	1772	hemileia vastatrix	497
3	arabica	1676	coffea arabica	498
4	catuaí	1203	of coffee	452

Continua...

[10] Disponível em: <https://www.etermos.cnptia.embrapa.br/index.php>.
[11] Disponível em: < https://the.sketchengine.co.uk/login/>.
[12] Disponível em: <http://www.yworks.com/products/yed>.
[13] Disponível em: <http://sobek.ufrgs.br/>.
[14] Disponível em: <http://www.vosviewer.com/>.

Tabela 1. Continuação.

5	meloidogyne	1105	coffea canephora	354
6	coffee	2469	key words	347
7	vastatrix	988	catuaí vermelho	316
8	exigua	856	híbrido de timor	311
9	canephora	749	embrapa café	306
10	cercosporiose	685	leucoptera coffeella	269
11	paranaensis	653	e discussão	308
12	zambolim	626	resultados e discussão	298
13	hemileia	628	resultados e	305
14	hampei	610	material e métodos	300
15	colletotrichum	613	referências bibliográficas	392
16	were	1073	material e	311
17	nematóide	928	simpósio de pesquisa	242
18	coffeella	532	cercospora coffeicola	231
19	ufv	1006	et al	3276
20	ferrugem	2241	e métodos	314
21	plants	608	pesquisas cafeeiras	221
22	rust	543	hemileia vastatrix berk	217
23	progênies	551	café conilon	255
24	hypothenemus	470	congresso brasileiro de pesquisas	215
25	resistance	638	brasileiro de pesquisas cafeeiras	213
26	was	1254	leaf rust	208
27	gloeosporioides	442	mundo novo	455
28	plant	695	meloidogyne paranaensis	200
29	fastidioso	591	federal de lavras	276
30	conilon	546	inimigos naturais	281
31	iapar	540	colletotrichum spp	199
32	coffeicola	369	xylella fastidiosa	229
33	iac	972	universidade federal de lavras	264
34	nematode	364	bicho mineiro	193
35	catucaí	362	plantas de café	194
36	incognita	402	coffee leaf	190
37	fitopatologia	475	coffee berry	189
38	leucoptera	360	vigor vegetativo	184
39	ipr	452	curva de progresso	175
40	inoculação	557	brasileiro de pesquisas	221
41	cochonilha	441	meloidogyne spp	166
42	leaf	520	vastatrix berk	163
43	inocular	581	consórcio brasileiro	173
44	phoma	332	brasileiro de	206
45	progénie	311	produção de café	226
46	berk	303	congresso brasileiro de	169
47	cercospora	299	coffee plants	159
48	broca	723	vermelho iac	157
49	disease	413	métodos o	160
50	conídio	301	control of	164

Desse corpus, foram organizadas duas bases de dados discriminadas a seguir:

Base de Dados Terminológicos sobre doenças e pragas do cafeeiro – Módulo Corpus Textual

Esta base de dados está construída e disponível no aplicativo web e-Termos , operacional nos servidores da Embrapa Informática Agropecuária, cujo acesso está sujeito à autorização de seus administradores. A base é o conjunto de textos originados de artigos técnico-científicos (306 documentos; 637.067 palavras) sobre o domínio de conhecimento em questão, reunidos, limpos, compilados e preparados para extração semiautomática de candidatos a termos para geração de recursos terminológicos e semânticos de organização e representação da informação e do conhecimento.

Base de Dados Terminológicos sobre doenças e pragas do cafeeiro – Módulo Vocabulário

Esta base de dados está construída e disponível no aplicativo web e-Termos[15], operacional nos servidores da Embrapa Informática Agropecuária, cujo acesso está sujeito à utorização de seus administradores. A base é conjunto de termos (1295 palavras ou expressões) sobre o domínio de conhecimento em questão validados por especialistas para geração de recursos terminológicos e semânticos de organização e representação da informação e do conhecimento.

3.2. Mapeamento dos dados terminológicos sobre pragas e doenças do cafeeiro nos tesauros agropecuários Thesagro e Agrovoc

Nesta proposta de representação do conhecimento sobre pragas e doenças do cafeeiro, o mapeamento da terminologia extraída do corpus nos tesauros agrícolas revela um novo sistema de conceitos onde são indicados os rótulos semânticos que qualificam as relações entre os conceitos comuns compartilhados pelos três vocabulários (Figura 3, A). Dessa forma, as terminologias podem assumir a qualificação de um vocabulário controlado sobre o domínio e pode ser utilizada em processos de indexação documental e de recuperação de informação.

Além disso, a estrutura conceitual resultante desse mapeamento, quando salva em formato de arquivo *.GRAPHML permite que cada um dos conceitos possa ser explorado separadamente, gerando subsistemas como o exemplificado na Figura 3, B, para o termo “ferrugem”.

A importação dos dados dos três vocabulários e sua representação por meio dos recursos do software yEd, traz dois aportes significativos: a) a visualização do sistema de conceitos como um todo (Figura 3, A) ou de recortes específicos para cada termo (Figura 3, B), favorece a cognição do posicionamento do conceito no sistema geral; e b) de suas relações com os outros conceitos permitindo, ainda, que se edite tantos os nós como as arestas de maneira facilitada, agregando facilidades de gestão terminológica e semântica ao sistema de conceitos.

15 Disponível em: <http://www.etermos.cnptia.embrapa.br/>.

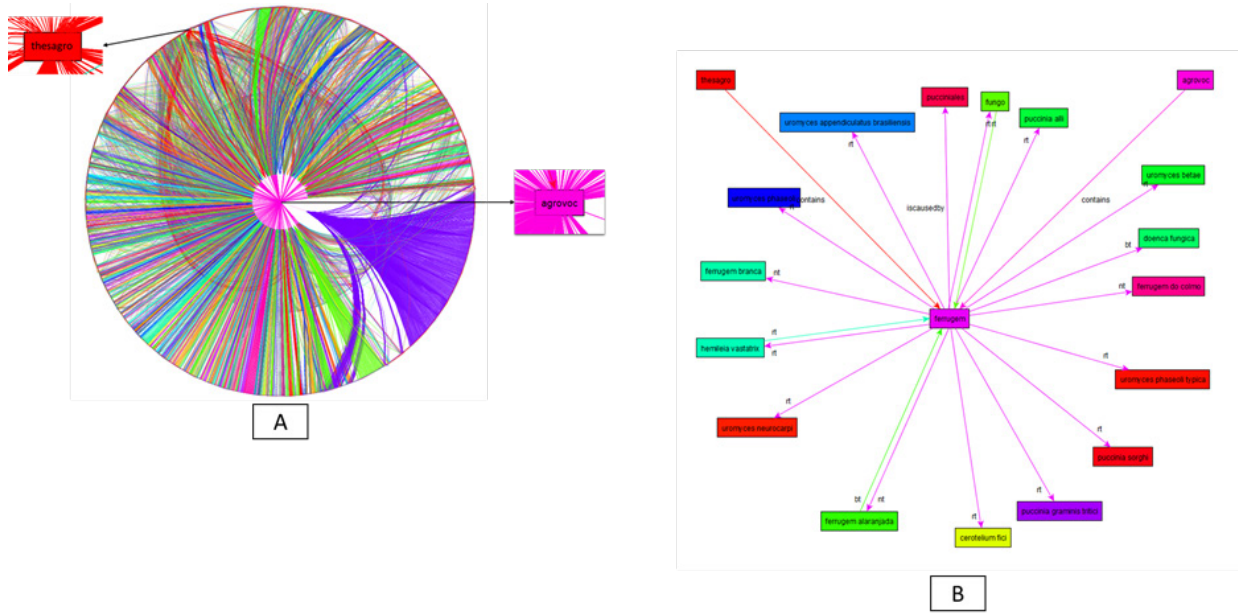

Figura 3. Visualização do sistema de conceitos gerado a partir do mapeamento da terminologia extraída automaticamente a partir do corpus textual sobre pragas e doenças do cafeeiro com os tesauros Thesagro e Agrovoc. A: visão geral; B. detalhe dos relacionamentos semânticos para o termo “ferrugem”. Software yEd, layout radial.

3.3. Esquema conceitual genérico para representação do domínio sobre pragas e doenças do cafeeiro no contexto do Manejo Integrado de Pragas (MIP)

Esse esquema conceitual (Figura 4) propõe uma representação do conhecimento do domínio sobre pragas e doenças do algodoeiro, partindo de seus elementos (ou conceitos) fundamentais: “pragas”, “doenças” e “cafeeiro”. A partir de cada um desses conceitos desdobram-se reciprocamente outros elementos/conceitos, configurando tanto as relações hierárquicas como as relações associativas compondo uma visão geral do domínio a partir de seus conceitos fundacionais.

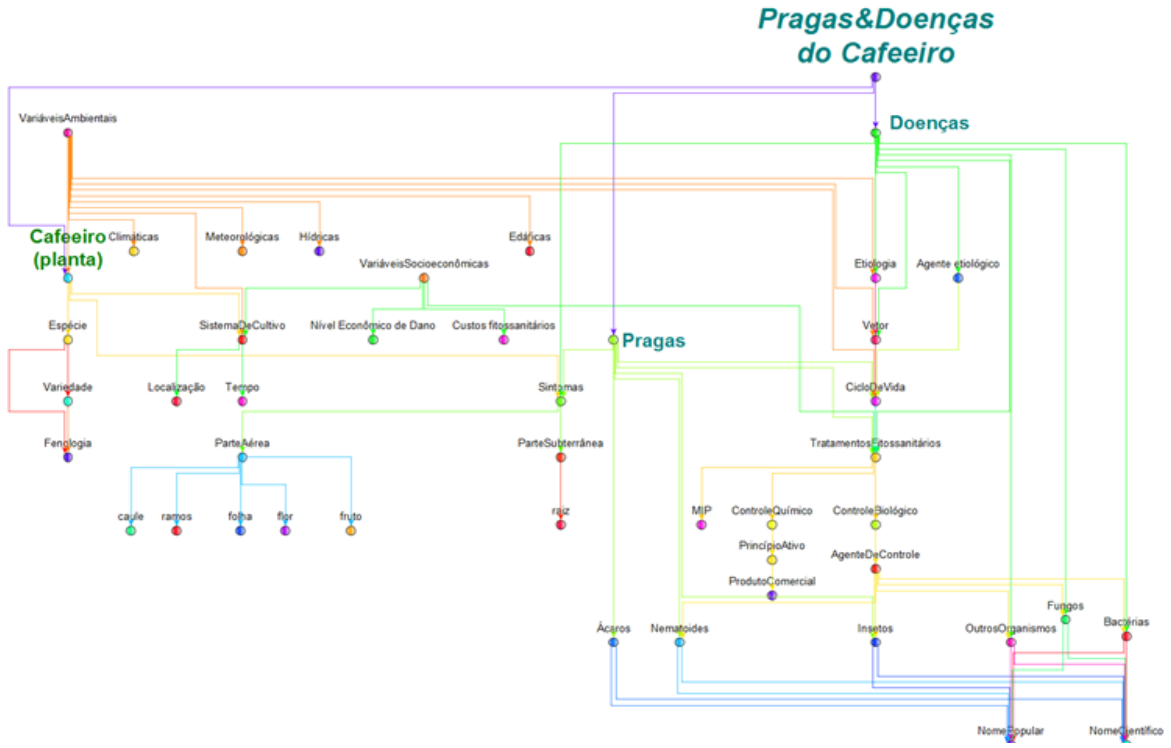

Figura 4. Esquema conceitual genérico (mapa mental) de representação livre do conhecimento sobre pragas e doenças do cafeeiro, reunindo e inter-relacionando alguns conceitos fundacionais que contextualizam entidades, objetos, processos e fenômenos envolvidos nesse domínio de conhecimento.

Esse mapa mental, sem compromisso de ser exaustivo e completo, organiza em primeira ordem os três elementos fundamentais a serem representados: 1) as doenças; 2) as pragas; e 3) a planta, no caso, o cafeeiro. A partir desses três nós, desdobram-se relações com outros conceitos, identificando-se cadeias hierárquicas lógicas, tais como exemplificadas nas Figuras 5, 6 e 7.

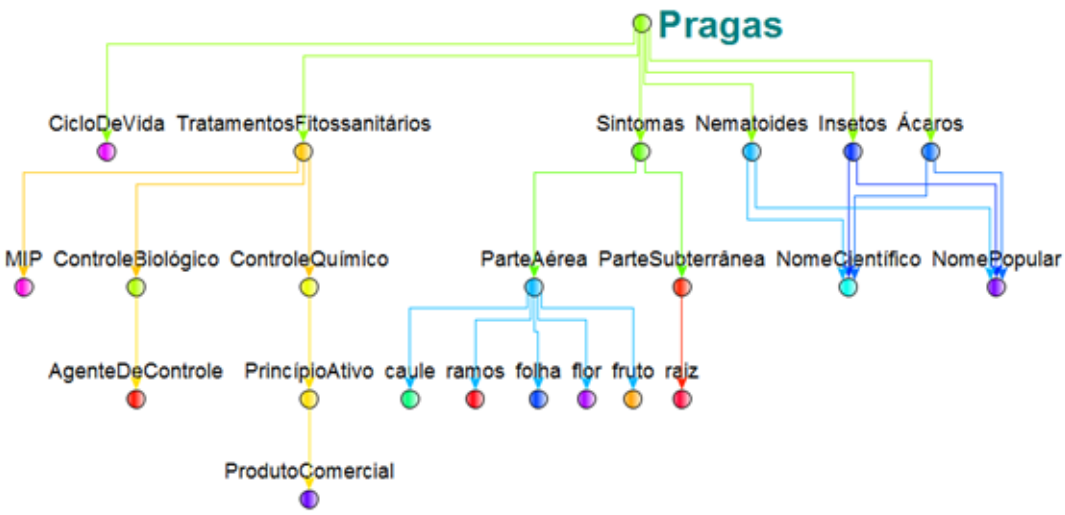

Figura 5. Recorte do mapa mental de representação do conhecimento sobre pragas e doenças do cafeeiro, ressaltando o desdobramento a partir do conceito “pragas”.

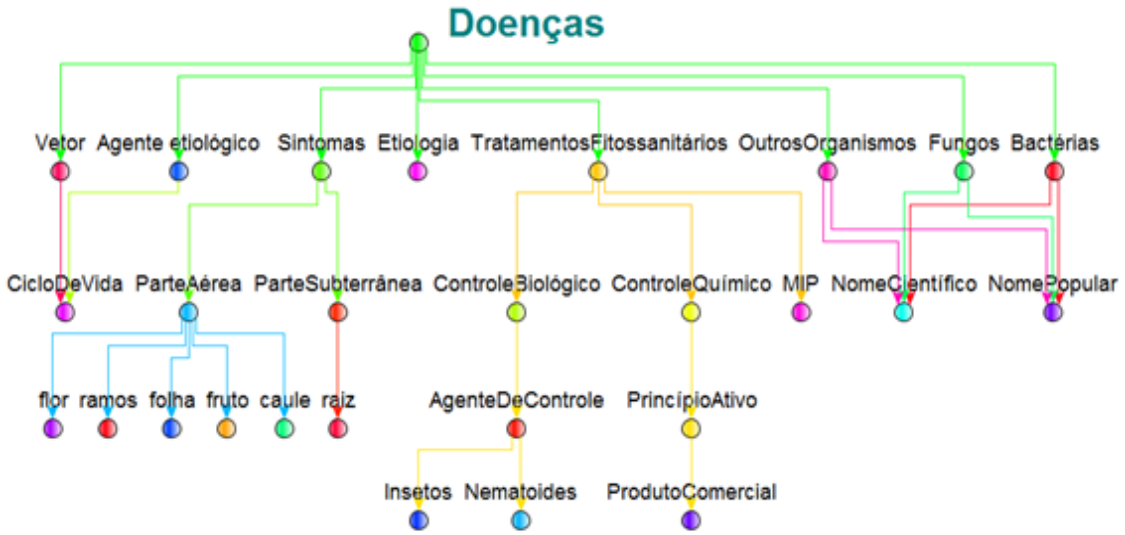

Figura 6. Recorte do mapa mental de representação do conhecimento sobre pragas e doenças do cafeeiro, ressaltando o desdobramento a partir do conceito “doenças”.

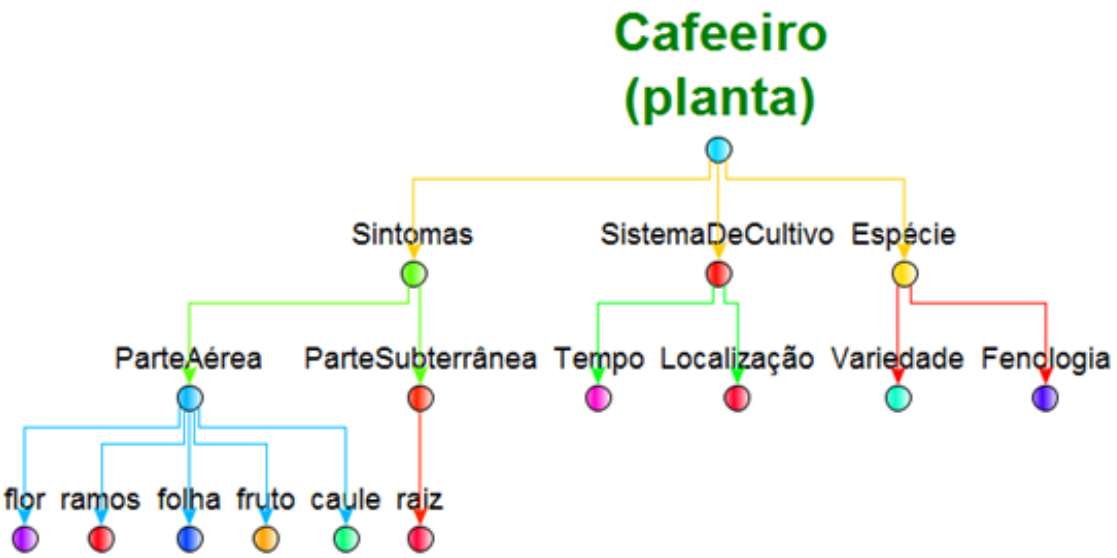

Figura 7. Recorte do mapa mental de representação do conhecimento sobre pragas e doenças do cafeeiro, ressaltando o desdobramento a partir do conceito “cafeeiro”.

As cadeias hierárquicas lógicas expressam relações conceituais que ao serem consideradas em conjunto constroem uma representação geral do domínio modelado ressaltando pontos de importância a serem considerados no entendimento ou compreensão geral do domínio modelado. Por exemplo: no monitoramento e manejo de uma determinada praga (Figura 5) é importante conhecer o ciclo de vida da praga, os sintomas do seu ataque que podem ser identificados em uma ou mais partes da planta. No caso das doenças (Figura 6), talvez seja importante identificar se existe um organismo vetor e seu próprio ciclo de vida, para contextualizar essa informação no processo de MIP. Ainda, no mesmo processo de MIP, talvez seja importante considerar o tipo de sistema de cultivo, assim como a espécie de café, a variedade e sua fenologia (Figura 7). Esse mapa deve ser entendido e utilizado como uma ferramenta dinâmica, que pode ser ajustado contínua e sistematicamente dependendo da necessidade de entendimento e exploração do domínio, com inclusão,

exclusão ou ajustes dos seus elementos (nós) ou de suas relações recíprocas. O próprio software escolhido para a edição do mapa permite essa plasticidade, tornando o mapa uma ferramenta operacional, tanto para fins cognitivos, como para base de modelagem computacional de sistemas de informação.

Esse esquema conceitual foi concebido e construído livremente, por especialistas em Ecologia Aplicada e Manejo Integrado de Pragas, com o intuito de ancorar os conceitos a serem minerados do corpus textual, esses sim, representativos da realidade da pesquisa, do desenvolvimento e da inovação praticados no âmbito desse domínio do conhecimento. Tal realidade é assegurada pelo processo de mineração de dados diretamente de corpus textual, compilado de artigos técnico-científicos apresentados num período de 13 anos sobre a temática em questão e que está representada pelo esquema conceitual, apresentado e explicado na seção 3.2.

3.4. Esquema conceitual resultante de mineração de dados em corpus textual compilado de artigos técnico-científicos sobre pragas e doenças do cafeeiro e Manejo Integrado de Pragas (MIP)

A Figura 8 representa um esquema ou mapa conceitual, construído por meio de técnicas de mineração de texto e de métricas estatísticas de frequência e coocorrência de palavras e expressões candidatas a termos presentes em um corpus textual. Esse procedimento reorganiza e representa os termos do texto num esquema relacional (rede) que evidencia hierarquias, equivalências e relações associativas dos termos (denotativos dos conceitos) pertinentes ao domínio de conhecimento a ser representado.

Além disso, tal esquema conceitual pode ser submetido às métricas convencionais utilizadas em Análise de Redes Sociais (ARS), uma vez que tanto quanto pessoas representam nós em uma rede social, conceitos/termos representam nós numa rede de conhecimento. Dessa forma, a relevância ou posicionamento de um determinado conceito, tomado como unidade de conhecimento, podem ser mensurados, revelando o papel daquele conceito na rede representativa do domínio, no caso "pragas e doenças do cafeeiro". Essas métricas, tais como: centralidade, grau, intermediação, proximidade, etc., podem fornecer informações sobre a relevância de entidades, competências, instituições, linhas de pesquisa, biodiversidade, métodos de manejo, localizações, etc., que se manifestam importantes na representação e entendimento do domínio de conhecimento modelado.

Os resultados dessa abordagem de representação revelaram o esquema conceitual apresentado na Figura 8, onde o tamanho dos nós (círculos) representa o peso ponderado de ligações que um determinado nó apresenta na rede, em outras palavras, representa a relevância do conceito/termo dentro do corpus analisado ou ainda seu contexto terminológico relativo, lembrando que tal corpus foi formado pela reunião e análise de 552 artigos apresentados no Simpósio de Pesquisa dos Cafés do Brasil, no período de 2000 até 2013 que totaliza 1,2 milhão de palavras.

Percebe-se, então, a relevância do termo "café" em primeiro plano, seguido pelos termos "ferrugem"; "produção", "plantas" e assim por diante. Os nós que aparecem não conectados à rede, representam termos que não alcançaram um índice suficiente para que o algoritmo estatístico de mapeamento pudesse ter identificado sua relevância na rede em termos de frequência ou coocorrência, considerando o limite de corte nos cem primeiros termos classificados e lembrando, ainda, que tais índices são ponderados em relação ao termo mais frequente, no caso, "café". Esse limite de corte é arbitrário e foi estabelecido apenas para garantir que o mapa não fique imensamente populado de nós e a rede se torne visualmente inadequada para análise. No entanto, a ferramenta

de construção e edição do mapa conceitual é robusta o suficiente para suportar a visualização de milhares de nós, caso seja do interesse do usuário.

Outro ponto a ser considerado nesse mapeamento é o significado representado por cada um dos termos (nós) constituintes da rede. Para melhor compreensão, é preciso esclarecer a diferença entre “mapa mental” e “mapa conceitual” uma vez que, no caso do presente trabalho, tem-se usado o termo “mapa mental”. No mapa mental, as relações que se estabelecem entre os nós (representadas pelas linhas que ligam dois nós) não são “rotuladas”, ou seja, não adquirem um valor semântico, denotando uma qualificação da relação estabelecida entre os dois conceitos ligados. Esse valor semântico poderia ser, por exemplo [“cafeeiro” éAtacadoPor “ferrugem”] ou [“ferrugem” éDoençaDe “cafeeiro”]. Quando um mapa mental ganha rotulagem semântica nas relações estabelecidas entre seus nós, passamos a chamá-lo de “mapa conceitual”. No caso da Figura 8, o mapa foi elaborado automaticamente tomando-se como propriedade de relação entre os conceitos os índices estatísticos de suas frequências e coocorrências relativas no corpus textual, calculado automaticamente pelo algoritmo embutido no software utilizado para mineração do texto (Sobek).

Face a esse esclarecimento, justifica-se a razão de os termos como “planta”, “plantas” e “plant” ou “minas” e “Minas Gerais” ou ainda “broca” e “broca do café” terem sido mantidos como nós separados no mapa, embora representem o mesmo conceito.

A fase seguinte, no itinerário de evoluir a proposta de representação do conhecimento sobre pragas e doenças do cafeeiro consiste no exercício de alinhamento e a reorganização do esquema conceitual inicialmente proposto (Figura 4) com o esquema conceitual representado na Figura 8. Esse exercício produziu um esquema conceitual final onde os conceitos essenciais do domínio encontram os conceitos (termos) extraídos do corpus textual. O resultado é um terceiro esquema conceitual (Figura 9) que alinha a conceitualização genérica de fundamentação do domínio com os resultados e conhecimentos empíricos relatados nos artigos técnico-científicos que compõem o corpus textual. O exercício de alinhamento é explicado em detalhes na seção seguinte deste documento.

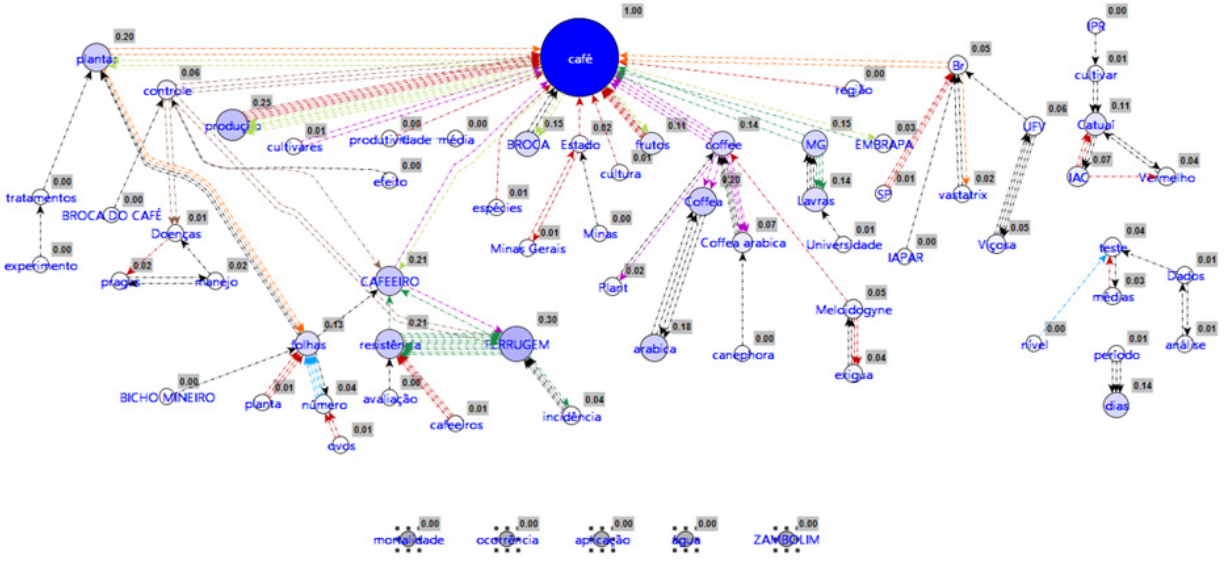

Figura 9. Esquema conceitual resultante de mineração de dados em corpus textual compilado de artigos técnico-científicos sobre pragas e doenças do cafeeiro e Manejo Integrado de Pragas (MIP).

3.5. Esquema conceitual resultante do alinhamento entre a proposta genérica e a proposta oriunda de mineração de dados textuais sobre pragas e doenças do cafeeiro e Manejo Integrado de Pragas (MIP)

Utilizando-se o software de edição e visualização de esquemas conceituais (yEd) os conceitos/termos dos dois esquemas conceituais anteriores (seções 3.3 e 3.4) foram reunidos e suas relações recíprocas redesenhadas, produzindo-se o mapa mental apresentado na Figura 10.

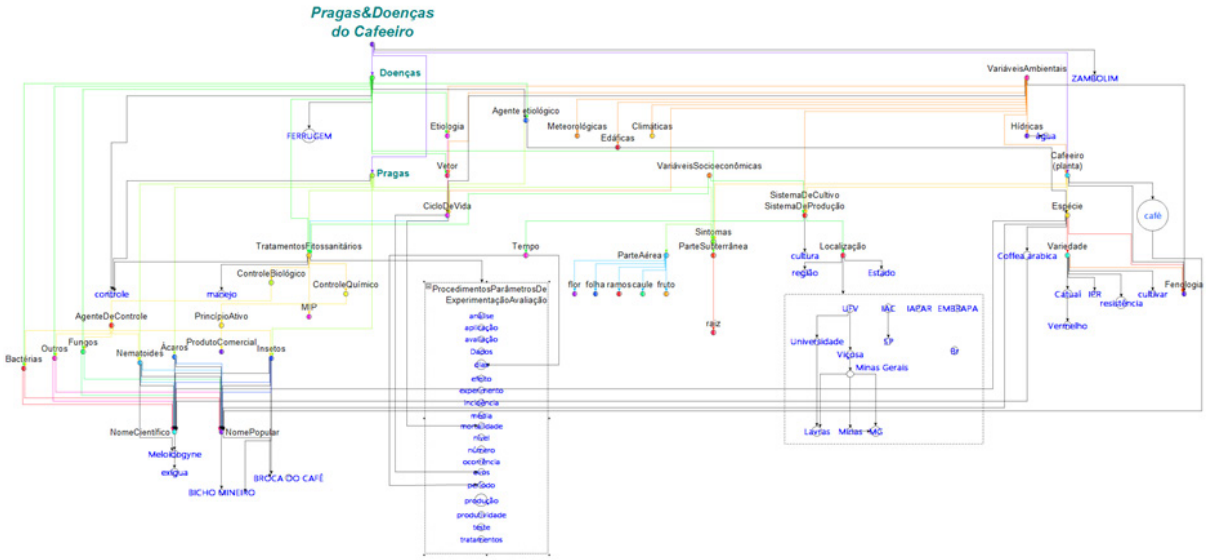

Figura 10. Esquema conceitual resultante do alinhamento entre a proposta genérica e a proposta oriunda de mineração de dados textuais sobre pragas e doenças do cafeeiro e Manejo Integrado de Pragas (MIP)

O esquema conceitual resultante demonstra que as facilidades de edição do software yEd permitem que esquemas conceituais construídos e editados separadamente possam ser agregados e alinhados, compondo uma nova proposta. Essa propriedade tecnológica viabiliza que os sistemas de conceitos sejam editados de maneira modular, tanto em estrutura como na composição dos fluxos de informações.

A propriedade também confere a vantagem de se construir os esquemas conceituais como “sistemas abertos”, uma vez que tendem a ser complexos em sua natureza: possuem inúmeros conceitos (nós) que mantém determinadas relações semânticas (arestas) em um contexto considerado - no caso, “pragas e doenças do cafeeiro e Manejo Integrado de Pragas (MIP)” mas que vão se manifestar de forma diferente se o contexto for outro, por exemplo, “impacto de mudanças climáticas sobre pragas e doenças do cafeeiro no Estado de Minas Gerais”.

Os recursos de edição e visualização do software yEd oferecem oportunidades de representação dos dados e da informação que favorecem a capacidade cognitiva e a inteligência decorrente da análise humana dos dados estruturados em esquemas conceituais. É uma facilidade ligada à capacidade de síntese desses recursos de representação do conhecimento que mantém o contexto (relações de causa-e-efeito) dos elementos componentes do sistema. Na prática, para um usuário de um sistema de informações concebido com base em tais esquemas, a informação e, portanto, o conhecimento decorrente, são mais facilmente apreendidos pois a percepção do significado da informação pode ser apreendida num processo de navegação pelo conteúdo informacional represen-

tado. Tecnologicamente, esses esquemas podem, por exemplo, dar base a conversão de conteúdos representados no formato de texto linear em hipertextos (Oliveira et al., 2016).

O que enriquece o valor de recursos de representação do conhecimento pela facilidade cognitiva que agrega no processo de "transferência" de conhecimento de um agente emissor para outro agente receptor.

3.6. Modelo de glossário sobre pragas e doenças do cafeeiro para veiculação em ambiente Web

A partir da análise do repertório de termos sobre pragas e doenças do cafeeiro, um modelo de glossário foi proposto visando a vinculação de um enunciado definitório, além de uma caracterização do termo mais formalizada do ponto de vista linguístico, gramatical e semântico permitindo assim, solução de problemas ligados à ambiguidade, apreensão cognitiva e comunicabilidade do termo no âmbito do sistema de conceitos e do domínio de conhecimento ao qual pertence.

Além da terminologia científica ou técnica, aproveitando a oportunidade da ocorrência relevante de vários nomes próprios (pessoas ou instituições), o modelo de glossário proposto inclui a possibilidade de apresentar tais entidades nomeadas igualmente com verbetes, agregando a elas informações pertinentes (nome completo, competências, endereço, siglas ou acrônimos), tornando-os parte integrante e complementar da proposta de representação do domínio do conhecimento sobre pragas e doenças do cafeeiro.

A Figura 11 (A, B) exemplifica fichas terminológicas preenchidas para discriminação de verbete técnico-científico e para entidade nomeada.

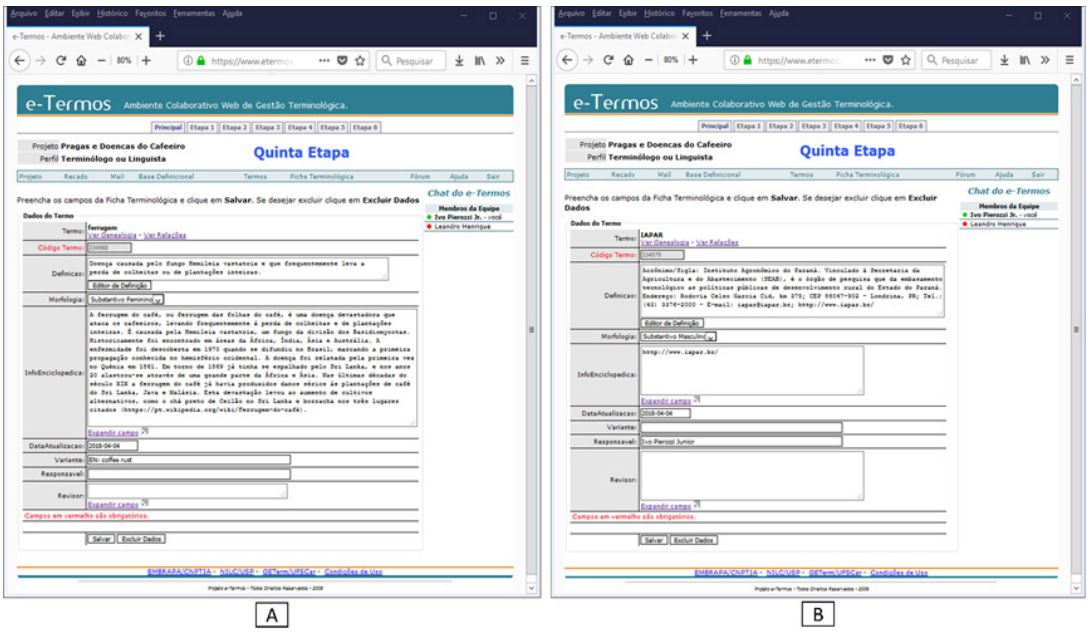

Figura 11. Modelos de fichas terminológicas para verbetes técnico-científico (A) e entidade nomeada (B), componentes do glossário sobre pragas e doenças do cafeeiro, construído no software e-Termos.

O modelo do glossário está publicado no endereço <https://www.etermos.cnptia.embrapa.br/produtos/produto330.html>. Para efeito de cumprimento dos objetivos do presente trabalho, apenas algumas fichas, referentes a poucos termos foram preenchidas com o intuito de exemplificarem como os verbetes podem ser trabalhados e publicados em ambiente web. O e-Termos, software utilizado para a concepção, construção e gestão do glossário é colaborativo. O modelo de glossário pode ser adotado por uma comunidade interessada em continuar seu desenvolvimento bastando, para isso, entrar em contato com a Embrapa Informática Agropecuária, por meio do Serviço de Atendimento ao Cidadão (SAC) da Embrapa[16].

3.7. Mapeamento semântico complementar sobre pragas e doenças do café: corpus textuais de bases de dados bibliográficos

1) Geração dos dados:

a) No. de documentos recuperados na busca:

b) SCO: 15; período: 1997-2017; número total de palavras no corpus: 5.682;

c) WoS: 33; período: 1996-2016; número total de palavras no corpus: 24.847;

2) Análise dos dados: parâmetros configurados no software VOSViewer para a visualização dos mapas:

a) SCO:

i) Número de termos incluídos no mapeamento semântico: 61 (considerando o critério do número mínimo de ocorrência de uma palavra = 1);

b) WoS:

ii) Número de termos incluídos no mapeamento semântico: 247 (considerando o critério do número mínimo de ocorrência de uma palavra = 1);

A Tabela 2 apresenta, respectivamente para as bases de dados bibliográficos Scopus e Web of Science, os 50 termos mais frequentes e coocorrentes extraídos no mapeamento semântico. O número de ordem dos termos está organizado pelo valor do “total link strength” (força de ligação total) que representa o número de vezes que um determinado termo se liga a outros termos da lista, o que coincide com o ranqueamento do valor de ocorrência absoluta do termo.

Tabela 2. As 50 palavras mais frequentes extraídas do corpus textual sobre pragas e doenças do cafeeiro originado em documentos recuperados das bases de dados Scopus (período: 1997-2017) e Web of Science (período: 1996-2016) com a expressão de busca: "coffee pest" OR "coffee diseases" AND Brazil.

	Scopus			Web of Science		
No.	keyword	occurrences	total link strength	keyword	occurrences	total link strength
1	coffea arabica	8	30	coffee pest	9	9
2	hemileia vastatrix	4	16	coffea arabica	6	6
3	brazil	2	15	coffee	6	6
4	coffee pest	2	15	hypothenemus hampei	6	6
5	coffee	2	10	coffee berry borer	3	3
6	geostatistics	1	10	coniothyrium	3	3
7	lepidoptera	1	10	pest control	3	3

Continua...

[16] Disponível em: <https://www.embrapa.br/fale-conosco/sac/>.

Tabela 2. Continuação.

8	lyonetiidae	1	10	spatial analysis	3	3
9	population monitoring	1	10	trap density	3	3
10	resistance	2	10	biological control	2	2
11	sampling range	1	10	coffee leaf miners	2	2
12	sex pheromone	1	10	ecosystem services	2	2
13	spatial analysis	1	10	hemileia vastatrix	2	2
14	trap density	1	10	lepidoptera	2	2
15	trap interaction	1	10	leucoptera coffeella	2	2
16	híbrido de timor	2	9	natural enemies	2	2
17	pest management	2	9	pathogen	2	2
18	coffee leaf rust	2	8	pest	2	2
19	Shypothenemus hampei Scoffea spp. Srianodine Scyazypyr™ Sbeneviа™	1	5	red spider mites	2	2
20	bacillus	1	5	sampling	2	2
21	bayesian analysis	1	5	bait	1	1
22	beneviа™	1	5	biotic and biotic habitat variables	1	1
23	biocontrol	1	5	acari	1	1
24	coffea spp.	1	5	agroecosystem	1	1
25	coffee leaf scorch	1	5	androctonus australis insect toxin	1	1
26	cyazypyr™	1	5	antagonism	1	1
27	defence	1	5	antestiopsis orbitalis	1	1
28	disease management	1	5	ants	1	1
29	genetic structure	1	5	attractant-baited traps	1	1
30	hypothenemus hampei	1	5	autonomous pest control	1	1
31	molecular phylogenetics	1	5	bayesian analysis	1	1
32	population differentiation	1	5	behavior	1	1
33	pseudomonas	1	5	berry borers	1	1
34	rianodine	1	5	biocontrol	1	1
35	silicon	1	5	biodiversity	1	1
36	c. dewevrei	1	4	bowman-birk inhibitor	1	1
37	c. kapakata	1	4	braconidae	1	1
38	c. racemosa	1	4	brazil	1	1
39	coffee leaf miner	1	4	caffeine	1	1
40	conservation biological control	1	4	coccus viridis	1	1
41	host plant	1	4	coffea canephora	1	1
42	leucoptera coffeella	1	4	coffea spp.	1	1
43	natural control	1	4	coffee agroecosystem	1	1
44	natural enemies	1	4	coffee berry borers	1	1
45	agricultural acarology	1	3	coffee farms	1	1
46	armaemors	1	3	competition	1	1
47	cercospora coffeicola	1	3	connectivity	1	1
48	coffea arabica l.	1	3	control	1	1
49	epidemiology	1	3	costa rica	1	1
50	green lacewing	1	3	development time	1	1

As Figuras 12 e 13 apresentam a visualização do mapeamento semântico complementar realizado a partir do corpus textual construído com dados de referências bibliográficas recuperadas das bases de dados Scopus e Web of Science, sobre o domínio de conhecimento sobre pragas e doenças do cafeeiro.

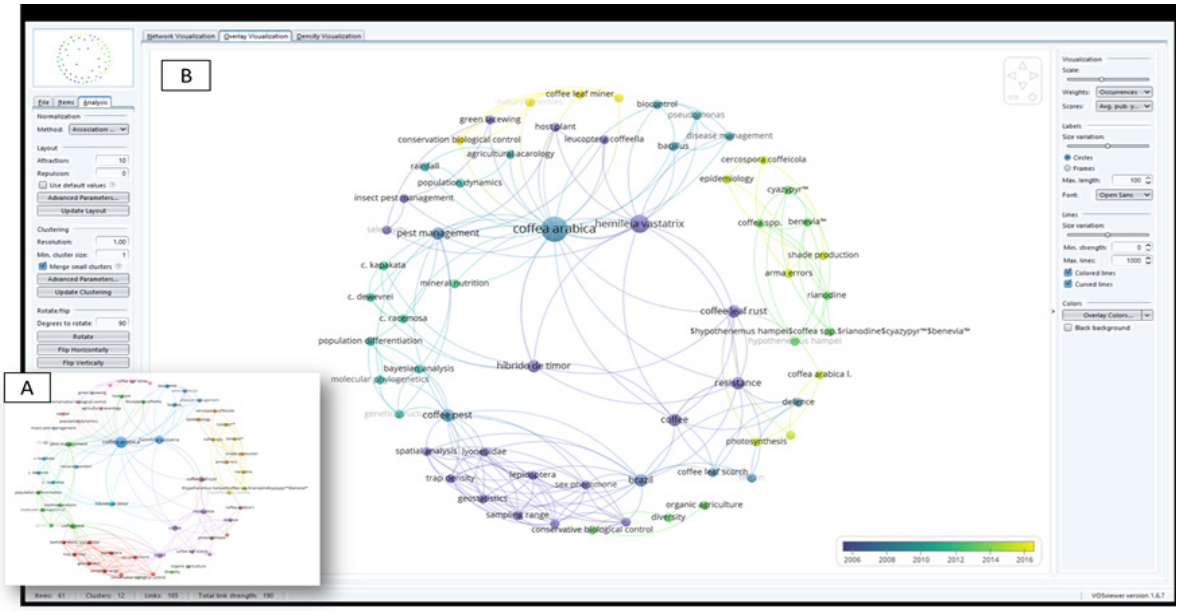

Figura 12. Mapeamento semântico: Base de dados Scopus; expressão de busca: "coffee pest" OR "coffee diseases" AND Brazil; 21 documentos; 61 termos; 12 agrupamentos. A: mapa geral; B: mapa de agrupamentos com sobreposição de visualização, onde cores frias representam termos presentes em documentos mais antigos e cores mais quentes representam termos presentes em documentos mais recentes.

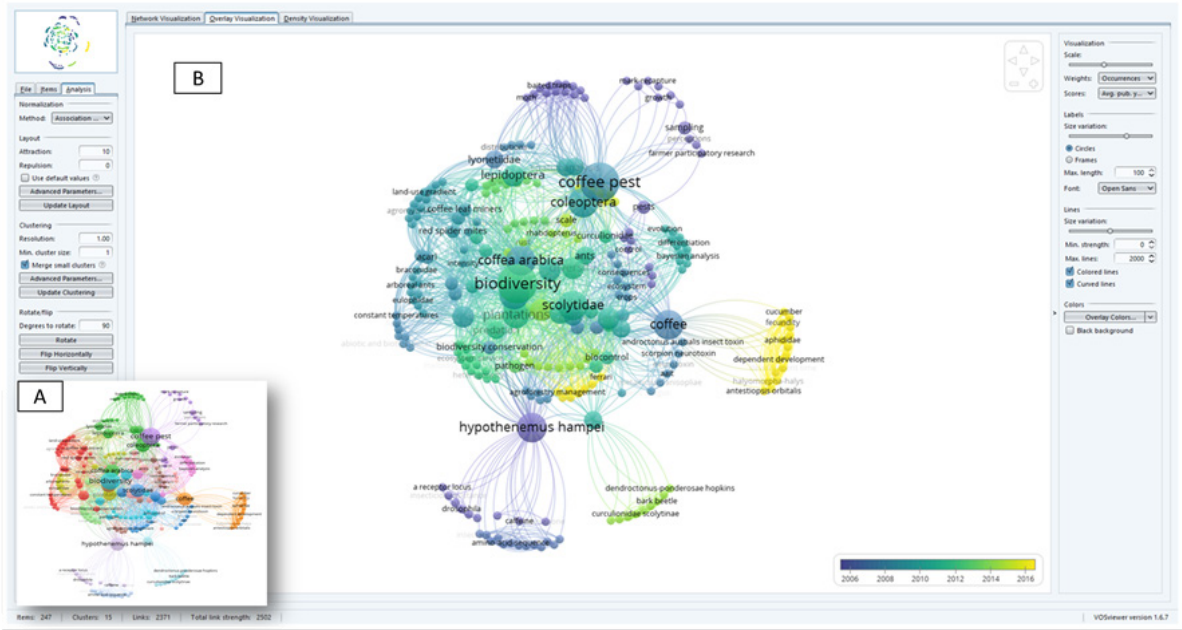

Figura 13. Mapeamento semântico: Base de dados Web of Science; expressão de busca: "coffee pest" OR "coffee diseases" AND Brazil; 33 documentos; 247 termos; 15 agrupamentos. A: mapa geral; B: mapa de agrupamentos com sobreposição de visualização, onde cores frias representam termos presentes em documentos mais antigos e cores mais quentes representam termos presentes em documentos mais recentes.

4. Considerações finais e perspectivas de evolução para os recursos de representação do conhecimento da temática sobre pragas e doenças do cafeeiro

Como o corpus textual de base, utilizado no presente trabalho para a representação do conhecimento, foi composto de documentos provenientes de um evento nacional bastante específico no contexto das pragas e doenças do cafeeiro, os recursos aqui desenvolvidos ficaram vinculados a uma representação extremamente fiel ao cenário do conhecimento brasileiro sobre o tema e explicitado em língua portuguesa.

Assim, um mapeamento semântico complementar foi realizado, dessa vez utilizando corpus textual em língua inglesa com a intenção de comparar os resultados originados de corpus representativos de cenários diferentes (nacional versus internacional). Para isso, considerando as especificidades do idioma inglês, foi utilizado um outro software (e, implicitamente, outro algoritmo de extração de automática de termos). O software VOSViewer foi então testado para esse fim, lembrando que além de extrator de termos o software também é um visualizador de dados, com a elaboração automática de mapas semânticos com base em frequência e coocorrência de termos em um corpus.

Com isso, além do tratamento linguístico/idiomático diferenciado, o software VOSViewer foi avaliado em relação às suas propriedades de geração de recursos de representação de conhecimento complementares: extração de termos; mapeamento semântico e visualização da informação. Nessa perspectiva, o VOSViewer se revela um software cujas funcionalidades complementam e agregam valor ao processo de representação do conhecimento quando for de interesse adicionar resultados em língua inglesa.

O exercício de representação do conhecimento aqui relatado agrega valor ao conjunto dos outros resultados produzidos no âmbito do projeto “Tecnologia da Informação para o manejo integrado de doenças e pragas do cafeeiro: modelagem, representação do conhecimento e ferramentas computacionais de diagnóstico e alerta” na medida em que os esquemas conceituais e recursos terminológicos desenvolvidos possam ser utilizados metodológica e computacionalmente como suporte:

- ao desenvolvimento de soluções para navegabilidade e acessibilidade em banco de dados específicos e na web;
- à compatibilidade e desambiguação terminológicas e à interoperabilidade semântica entre sistemas de informação relativos a essa temática;
- à implantação de sistemas de recuperação da informação;
- à realização de estudos de usuários para adequação de ferramentas e instrumentos de apoio à recuperação de informação na Embrapa

Referências

ALMEIDA, G. M. de B; CORREIA, M. Terminologia e *corpus*: relações, métodos e recursos. In: TAGNIN, S. E. O.; VALE, O. A. (Org.). **Avanços da linguística de *corpus* no Brasil.** São Paulo: Humanitas, 2008. p. 67-94.

CABRÉ, M. T. El principio de poliedricidad: la articulación de lo discursive, lo cognitive y lo lingüístico en Terminología (I). **Ibérica**, v. 16, p. 9-36, 2008.

CABRÉ, M. T. **La terminología**: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Linguística Aplicada/, 1999. 369 p. (Sèrie monografies, 3).

CABRÉ, M. T. **La terminología**: teoría, metodología, aplicaciones. Barcelona: Editorial Antártida: Empúries, 1993.

CABRÉ, M. T. La terminología hoy: concepciones, tendências y aplicaciones. **Ciência da Informação**, v. 24, n. 3, 1995.

DAHLBERG, I. Brief communication: what is Knowledge Organization? **Knowledge Organization**, v. 41, n. 1, p. 85-91, 2014. DOI: 10.5771/0943-7444-2014-1-85.

DAHLBERG, I. Teoria do conceito. **Ciência da Informação**, v. 7, n. 2, p. 101-107, 1978.

DI FELIPPO, A.; ALUÍSIO, S. M.; OLIVEIRA, L. H. M. de; ALMEIDA, G. M. B. de. OntoMethodus - a methodology to build domain-specific ontologies and its use in a system to support the generation of terminographic products. In: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM, 6., 2008, Vila Velha, ES. **Anais**... Espírito Santo: Universidade Federal do Espírito Santo, 2008. p. 393-395.

MACULAN, B. C. M. D. dos S. **Estudo e aplicação de metodologia para reengenharia de tesauro**: remodelagem do THESAGRO. 2015. 339 p. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais, Belo Horizonte.

NETTO, C. M.; LIMA, G. A. B. de O.; PIEROZZI JÚNIOR, I. An application of facet analysis theory and concept maps for faceted search in a domain ontology: preliminary studies. **Knowledge Organization**, v. 43, n. 4, p. 254-264, 2016.

OLIVÉ, A.; CABOT, J. A research agenda for conceptual schema-centric development. In: KROGSTIE, J.; OPDAHL, A. L.; BRINKKEMPER, S. (Ed.). **Conceptual modelling in information systems engineering**. Berlin: Springer, 2007. p. 319-334.

OLIVEIRA, E. D.; MACULAN, B. C. M. dos S.; PIEROZZI JUNIOR, I. Estruturação de hipertextos: proposta de conversão de textos. **Revista ACB**: Biblioteconomia em Santa Catarina, v. 21, n. 3, p. 564-578, ago./nov. 2016.

PIEROZZI JUNIOR, I.; SOUZA, M. I. F.; TORRES, T. Z.; OLIVEIRA, L. H. M. de; QUEIROS, L. R. Gestão da informação e do conhecimento. In: MASSRUHÁ, S. M. F. S.; LEITE, M. A. de A.; LUCHIARI JUNIOR, A.; ROMANI, L. A. S. (Ed.). **Tecnologias da informação e comunicação e suas relações com a agricultura**. Brasília, DF: Embrapa, 2014. Cap. 12. p. 235-280. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/126971/1/capitulo12-085-14.pdf>. Acesso em: 28 nov. 2019.

SARDINHA, T. B. Lingüística de Corpus: histórico e problemática. **DELTA**: Documentação de Estudos em Lingüística Teórica e Aplicada, v. 16, n. 2, p. 323-367, 2000. DOI: http://dx.doi.org/10.1590/S0102-44502000000200005.

SOUZA, R. R.; TUDHOPE, D.; ALMEIDA, M. B. A tentative typology of knowledge organization systems. In: INTERNATIONAL ISKO CONFERENCE, 11., 2010, Rome. **Paradigms and conceptual systems in knowledge organization**: proceedings. Würzburg: Ergon Verlag, 2010. p. 122-128. (Advances in knowledge organization, 12).

ZENG, M. L. Knowledge Organization Systems (KOS). **Knowledge Organization**, v. 35, n. 2-3, p. 160-182, 2008.

Embrapa
Informática Agropecuária

MINISTÉRIO DA
AGRICULTURA, PECUÁRIA
E ABASTECIMENTO

PÁTRIA AMADA
BRASIL
GOVERNO FEDERAL

CGPE 15657